Data Analysis Fundamentals Assignment

German Credit Data Analysis

Submitted By
Iftekhar Hossain

Data analysis analyze the raw data and convert it into meaningful information so that businesses can take more effective decision on their businesses and make more profit.
Banks are making profit on the loan they gave to different applicants. But loan defaulters cause loss to their business. Giving loan is always a risky business. So, the bank takes a decision when processing a loan application. Will they give loan to this applicant or not. They consider it as a Good Risk or a Bad risk.

Good Risk are those where the applicant is high likely to pay back the loan.
Bad Risk are those where the applicant is not likely to pay back the loan.

The dataset contains 10 columns and 1000 rows. Each row denotes an applicant who applied for a loan from the bank. Each applicant is classified either as a good credit risk or as a bad credit risk according to his/her values of the 09 features or variables as Age, Sex, Job, Housing, Saving accounts, Checking account, Credit amount, Duration and Purpose. Risk is the dependent variable, and the other nine variables are independent variables.

How many numeric variables are there?
Answer: There are four numeric variables as follows

- Age,
- Credit amount,
- Duration,
- Job

How many string variables are there?
Answer: There are five string variables as follows

- Sex (male, female),
- Housing (own, rent, free),
- Saving accounts (little, moderate, rich, quite rich),
- Checking account (little, moderate, rich),
- Risk (good or bad)

How many text variables are there?

Answer: There is one text variable as Purpose (radio/TV, education, furniture/equipment, car, business, domestic, appliances, repairs, vacation/others)

List of all the variables and about their data:

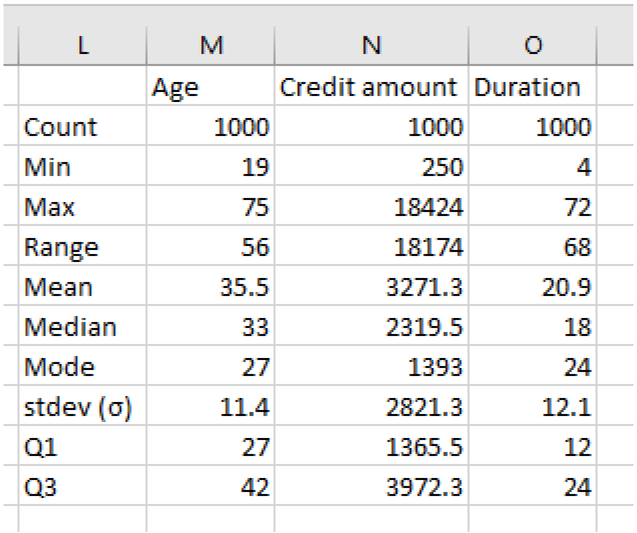
Nine Independent and One dependent Variables.

1. Age (numeric: in years)
2. Sex (string: male, female)
3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
4. Housing (string: own, free, rent)
5. Saving accounts (string - little, moderate, rich, quite rich)
6. Checking account (string)
7. Credit amount (numeric, in Deutsch Mark)
8. Duration (numeric, in months)
9. Purpose (text: (radio/TV, education, furniture/equipment, car, business, domestic, appliances, repairs, vacation/others)
10. Risk (string: good or bad) we can convert it to binary 0 (for Bad) or 1(for Good)

Number of unique values in those variables:

1. Age: 53
2. Sex: 02
3. Job: 04
4. Housing: 03
5. Saving accounts: 04
6. Checking account: 03
7. Credit amount: 921
8. Duration :33
9. Purpose: 08
10. Risk :02

Statistical Summary of the Data:

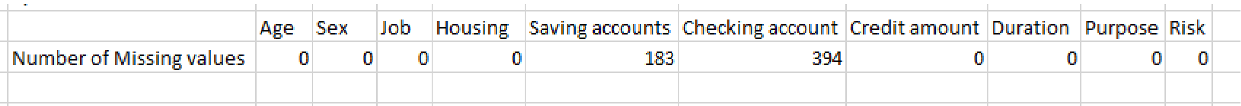


L	M	N	O
	Age	Credit amount	Duration
Count	1000	1000	1000
Min	19	250	4
Max	75	18424	72
Range	56	18174	68
Mean	35.5	3271.3	20.9
Median	33	2319.5	18
Mode	27	1393	24
stdev (σ)	11.4	2821.3	12.1
Q1	27	1365.5	12
Q3	42	3972.3	24

- in Age, Credit amount and in Duration the mean is greater than the median.

How many variables have missing values?
Answer: Two variables have missing values as ("NA").

1. Saving accounts
2. Checking account



	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
Number of Missing values	0	0	0	0	183	394	0	0	0	0

How many missing values are in each of the missing variables?
Answer:

1. Saving accounts (variable):_There are 183 missing values.

2. Checking account (variable):_There are 394 missing values.

What happens to the data if the records with missing data are removed?
Answer: Total numbers of missing values are 577 which are almost half of the data of 1000 observations. if we remove the records with the missing values, we shall lose significant amount of data.

The dataset has how many variables?
Answer: The dataset has ten variables.

Nine Independent and One dependent Variables.
a. Numeric; Age, Credit amount, Duration
b. Categorical; Sex, Job, Housing, Saving accounts, Checking account, Purpose
c. Binary: Risk

Age range. Answer: 56 years

The average age. Answer: 35.5 years

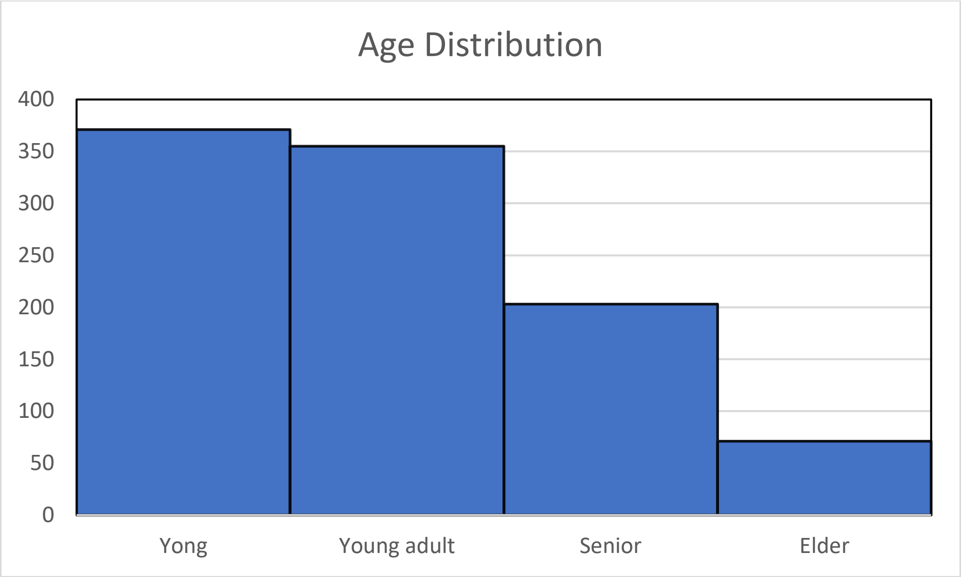
The average duration. Answer: 19 to 56 years

The range of credit. Answer: 250 to 18424

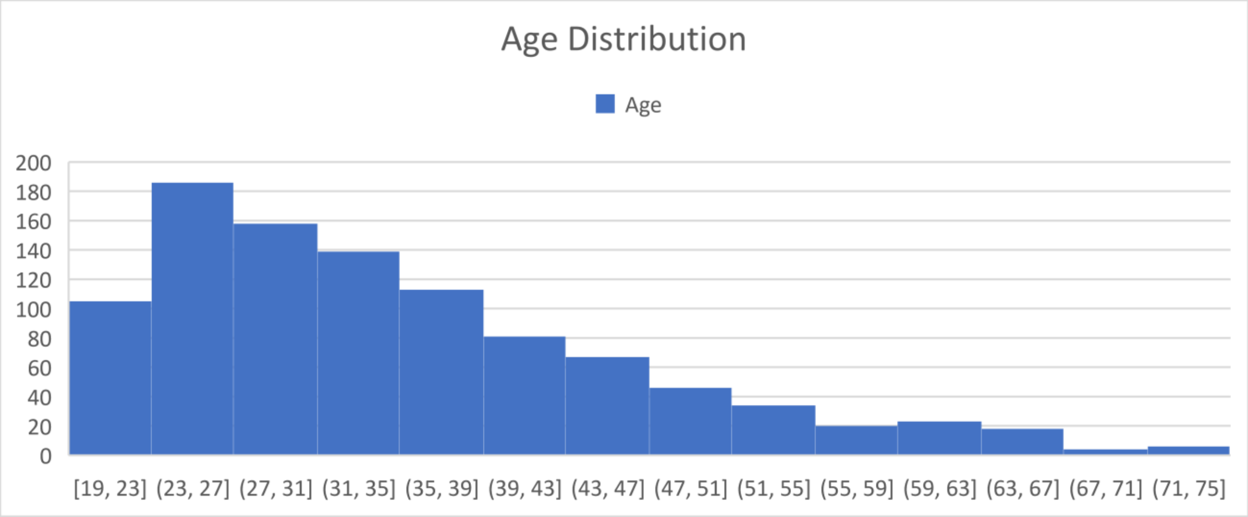
The average credit. Answer: 3271.30

Age Distribution:

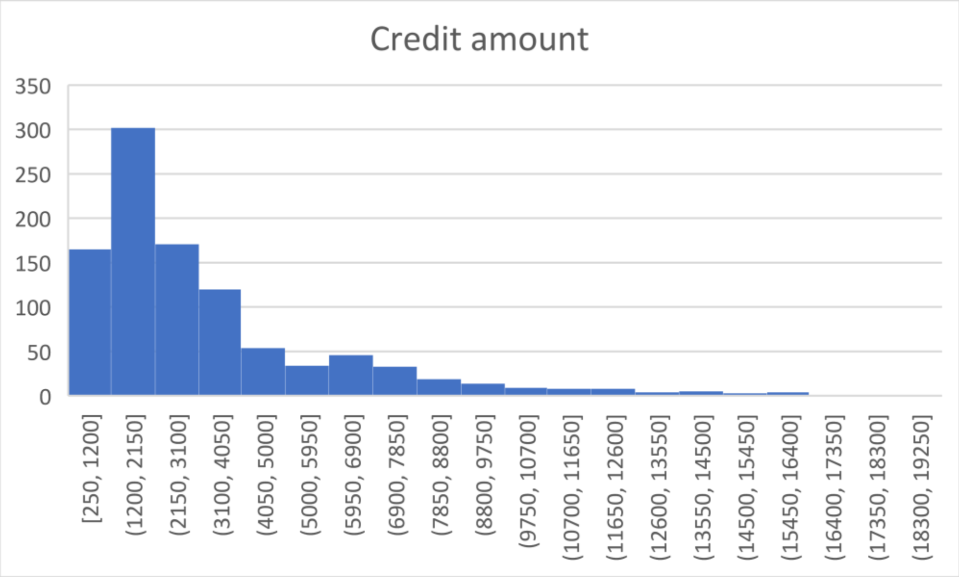
Bin	*Frequency*
Yong (19 -29)	371
Young adult (30 -40)	355
Senior (41-55)	203
Elder (55+)	71



- Yong (19 -29) people are more likely to apply for a loan and Elderly (55+) people are less likely to apply for a loan.

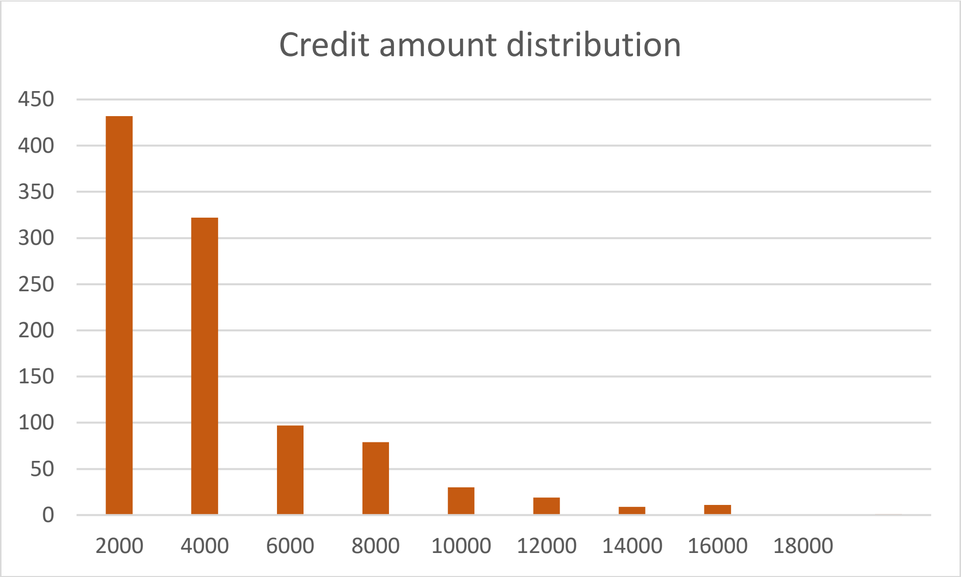


- The Age distribution is right-skewed.
- As we also seen before that in Age variable the mean is greater than the median.



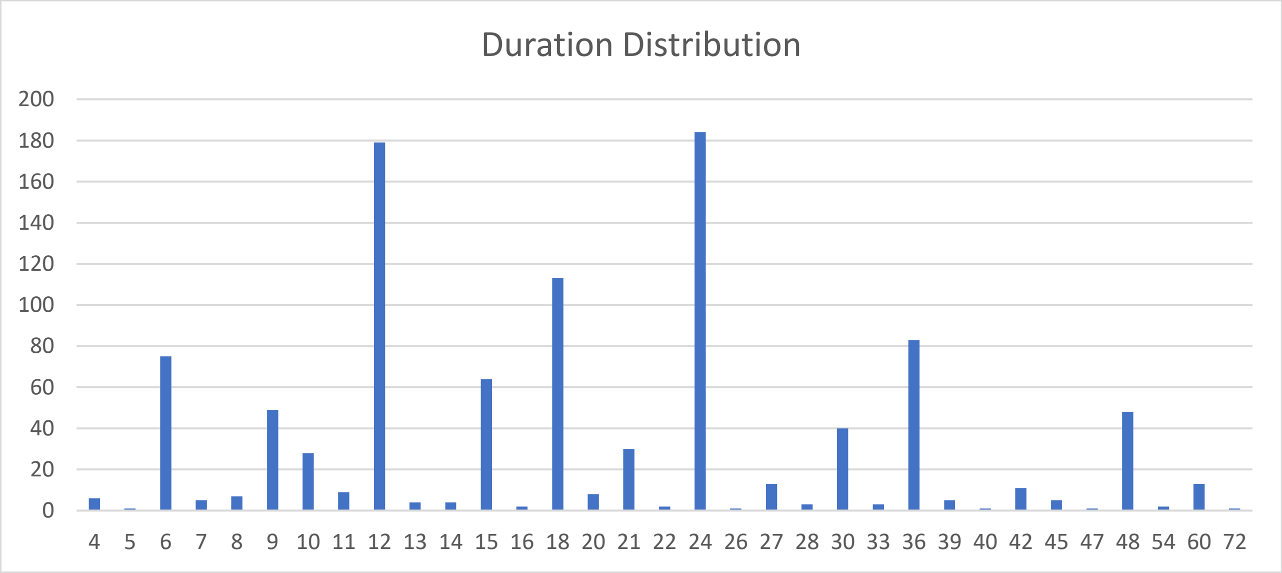
- The Credit amount distribution is right-skewed.
- As we also seen before that in Credit amount variable the mean is greater than the median.

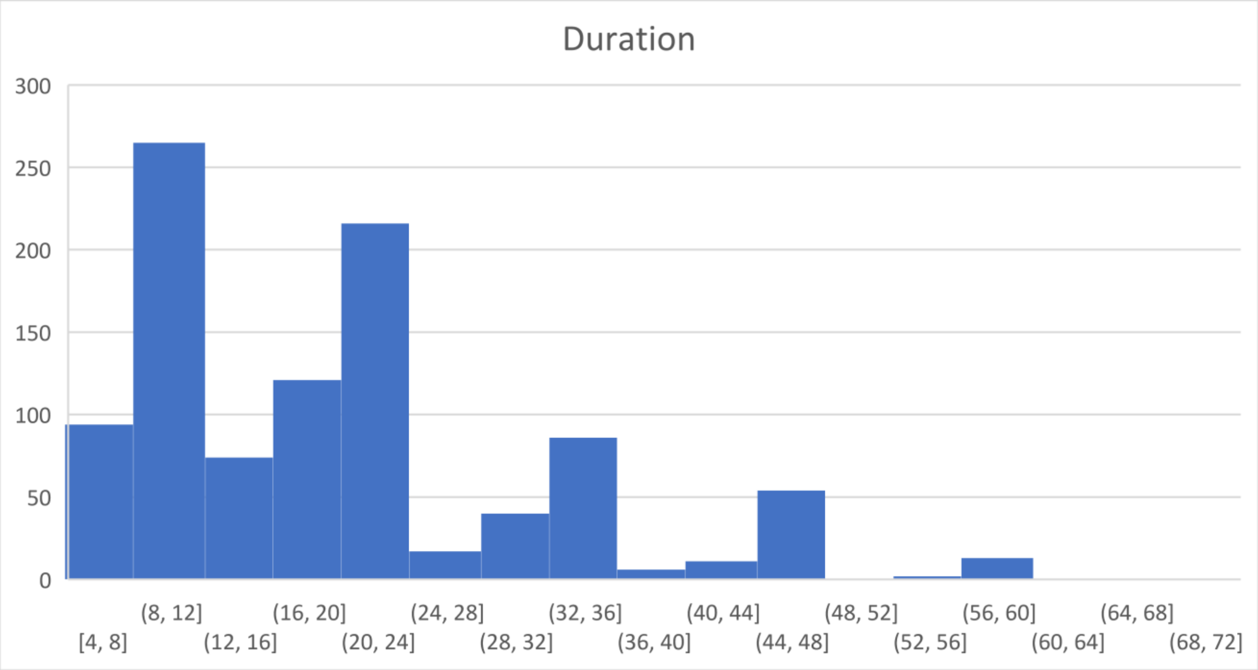
Credit amount (Bin)	Frequency
2000	432
4000	322
6000	97
8000	79
10000	30
12000	19
14000	9
16000	11
18000	0
18000+	1



- Number of applications for small amount is most likely higher than the application for larger amount.
- The Credit amount distribution is right-skewed.
- The applicants are most likely to apply for a loan of a small amount which is less than 4000 DM.

<u>Duration Distribution:</u>

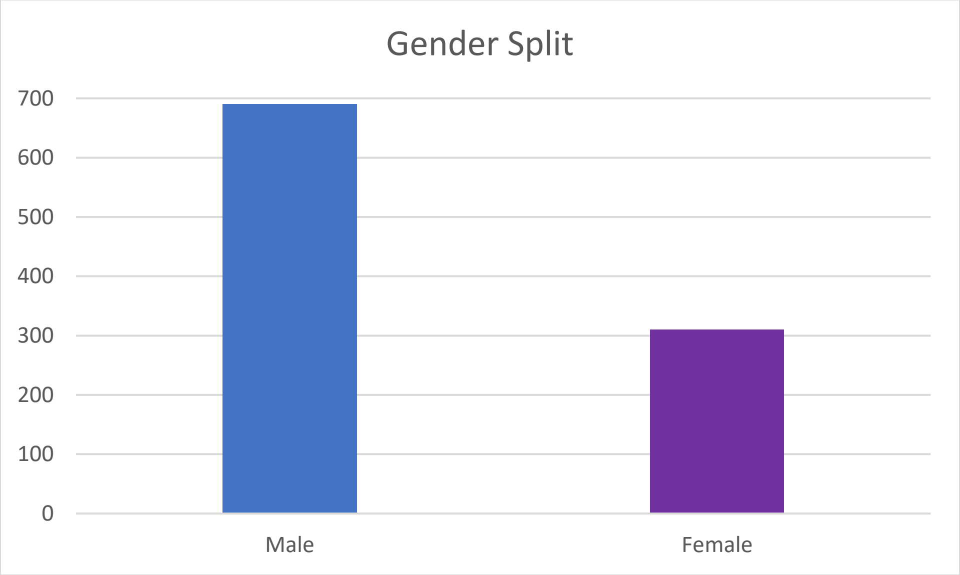


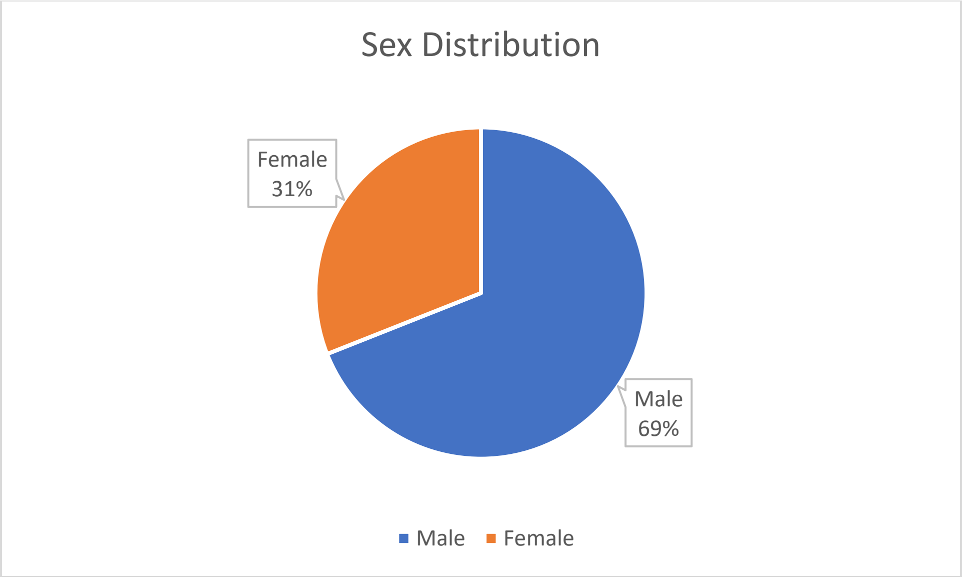


- The Duration distribution is right-skewed.
- Most of the loans have been paid off by 24 months.

Sex Distribution:

Male	690
Female	310



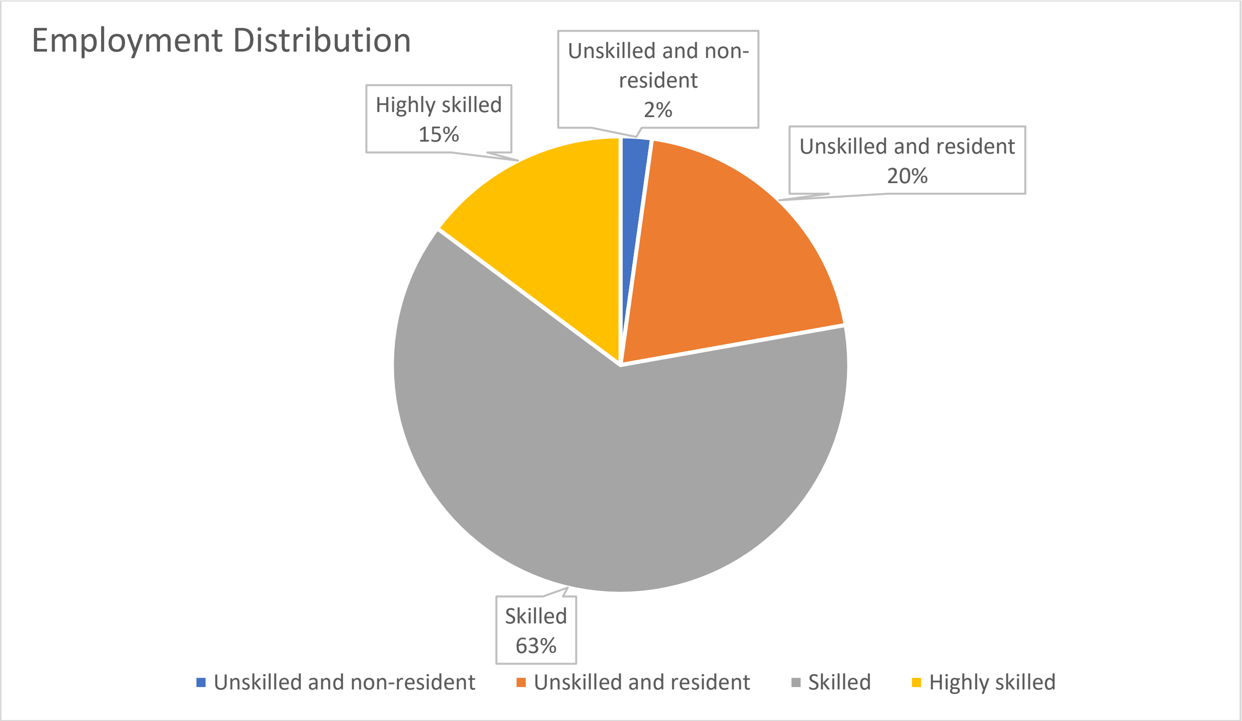


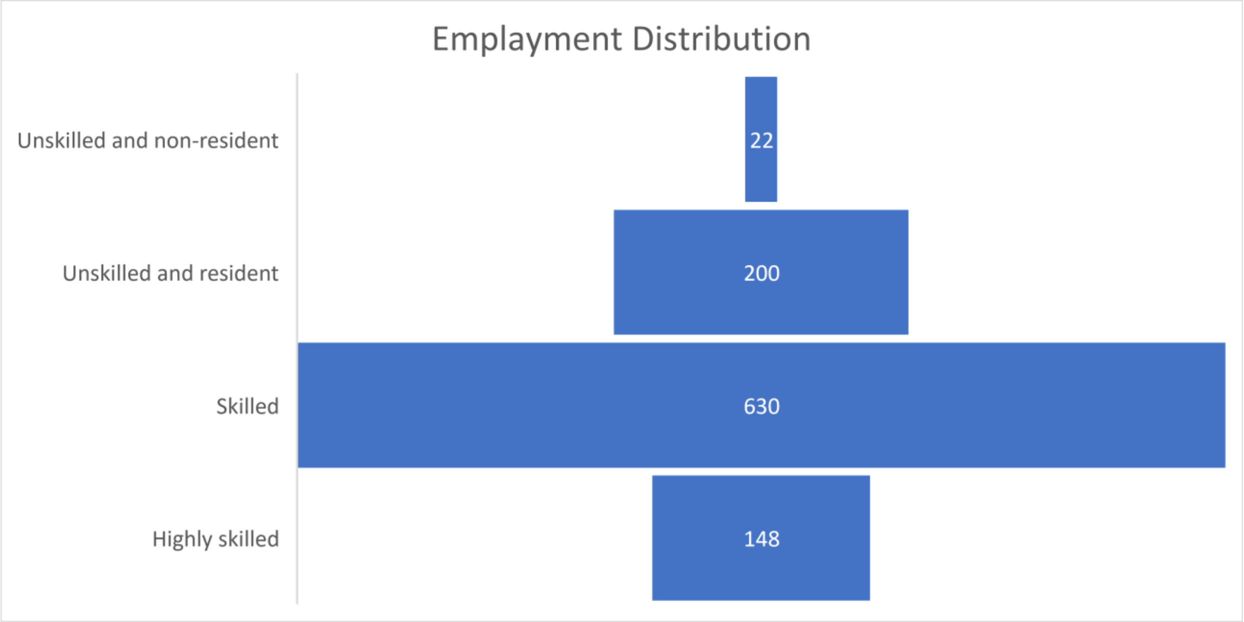
- Male applicants (690) are more than twice of the female applicant (310).
- Male applied for more loan application then Woman.

Job Distribution:

(0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)

Unskilled and non-resident	22
Unskilled and resident	200
Skilled	630
Highly skilled	148

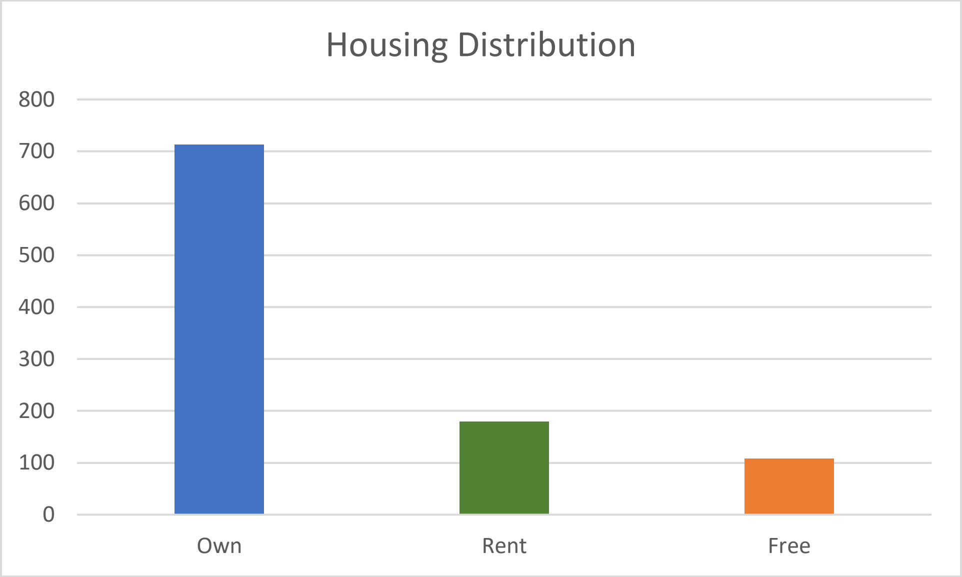




- Most of the applicants (63%) are from the skilled employment group.

Housing Distribution (own, free, rent):

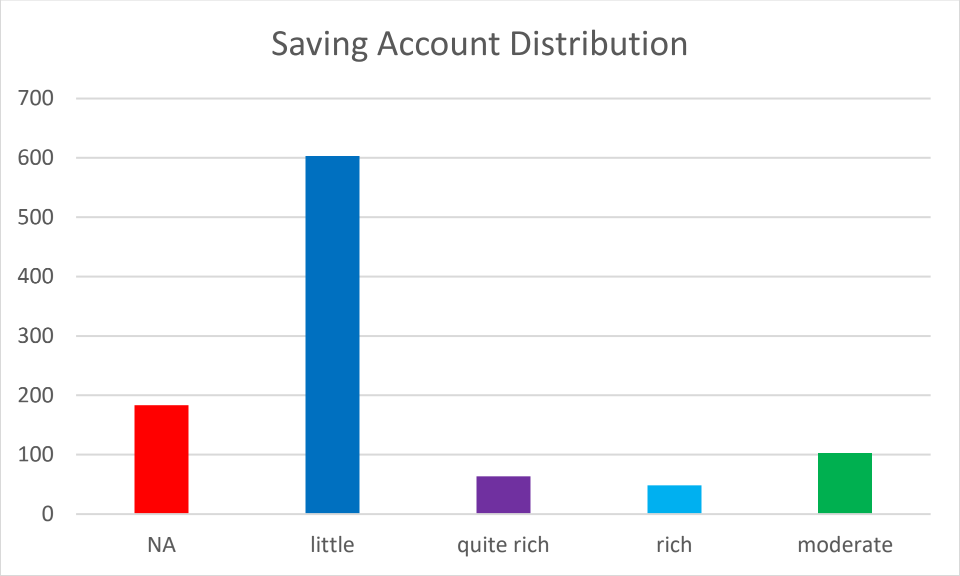
Own	713
Rent	179
Free	108

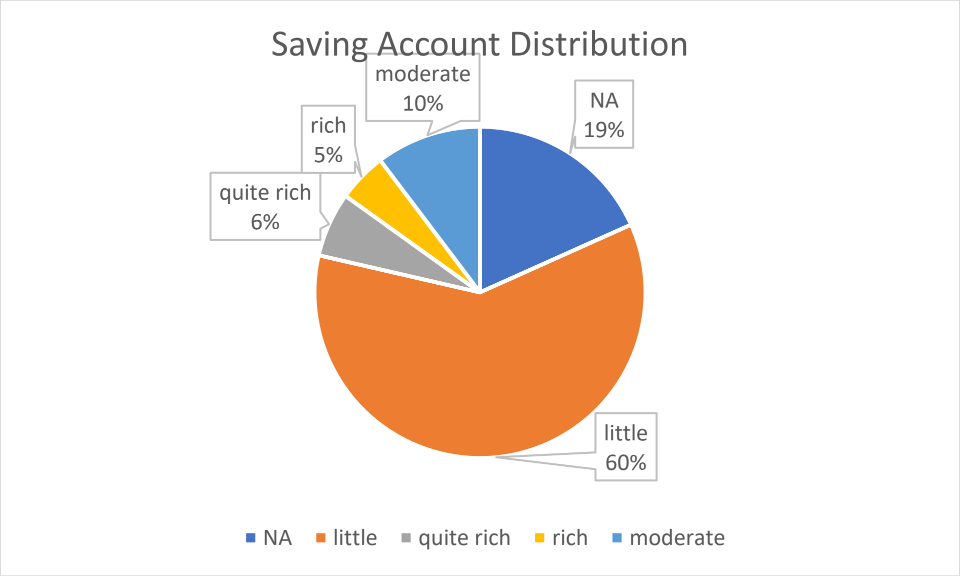


- Most of the applicants have their own house (713 out of 1000).

Saving accounts Distribution (little, moderate, rich, quite rich):

NA (Missing Values)	183
little	603
quite rich	63
rich	48
moderate	103

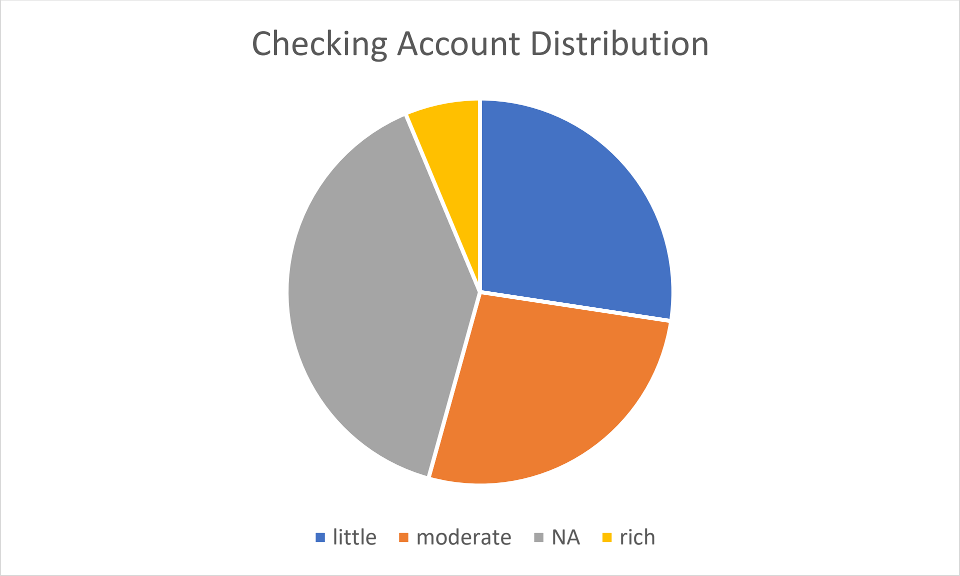




- In Saving accounts distribution, most of the applicant are from the **little** category.

Checking account Distribution:

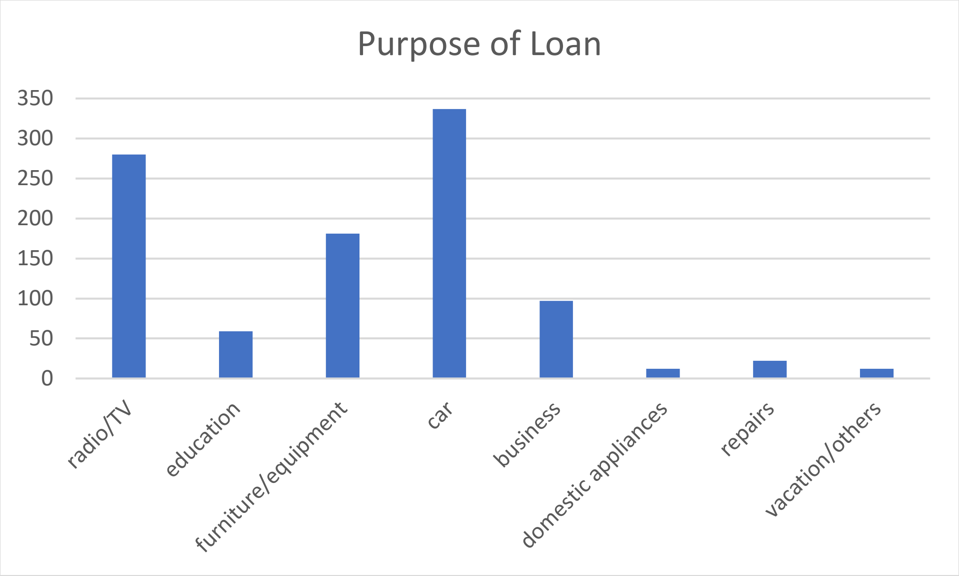
little	274
moderate	269
NA (Missing Values)	394
rich	63



- In Checking account Distribution, most of the applicants are from the **moderate** category.

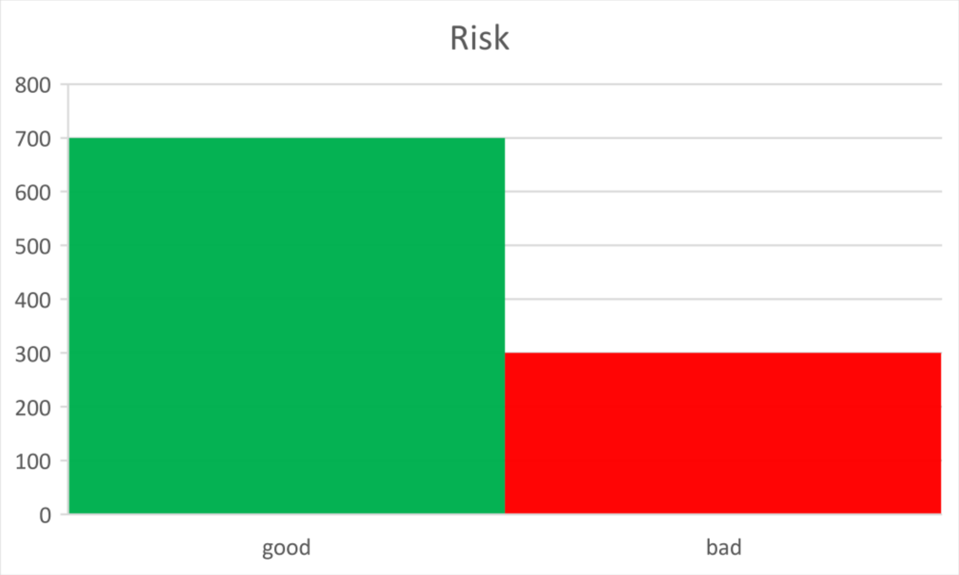
Purpose Distribution: (text: (radio/TV, education, furniture/equipment, car, business, domestic, appliances, repairs, vacation/others)

radio/TV	280
education	59
furniture/equipment	181
Car	337
business	97
domestic appliances	12
repairs	22
vacation/others	12



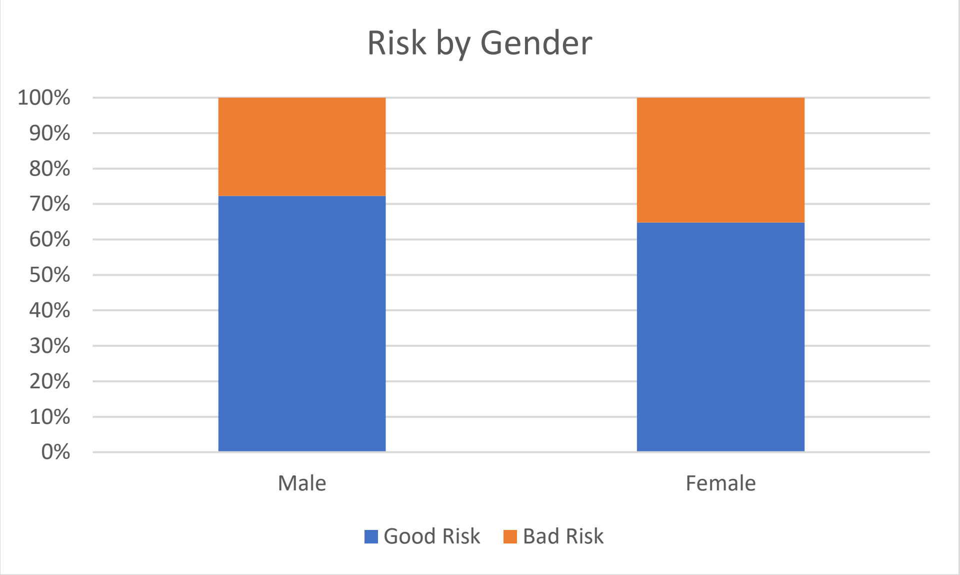
- Highest number of loans were applied for car 337. Second highest application number is for radio/TV.
- The lowest loan application is for domestic appliances 12 and vacation/others 12.
- We can see that most of the loans were taken for car and radio/TV.

Risk Distribution:



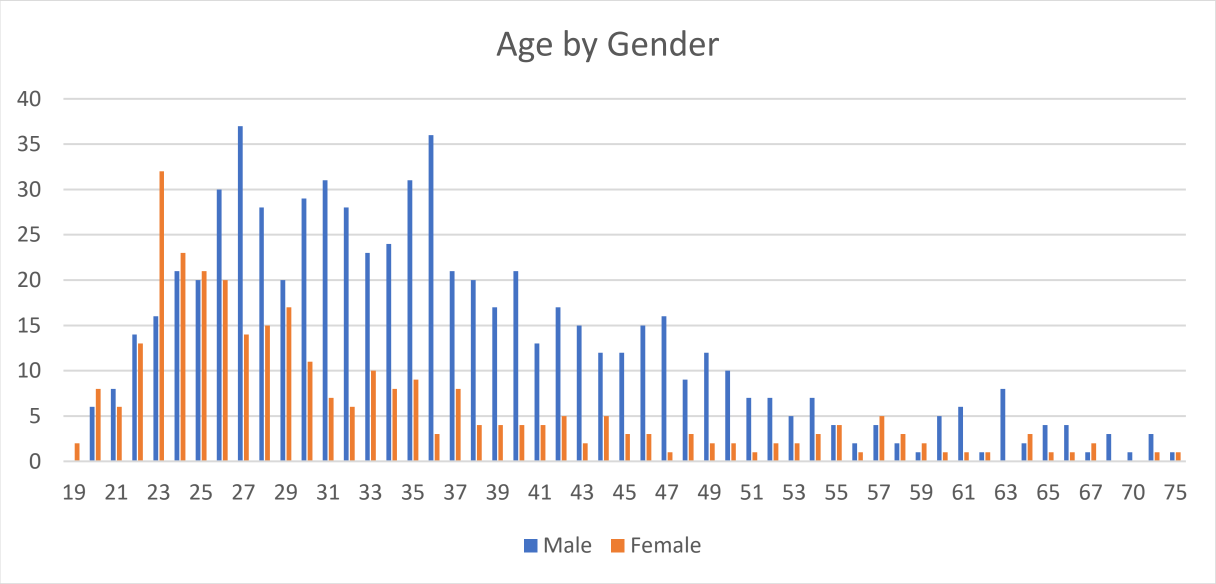
- 700 cases where the applicant was classified as **good** credit risk.
- 300 cases where the applicant was classified as **bad** credit risk.

Which gender has the better risk?
Answer: Man

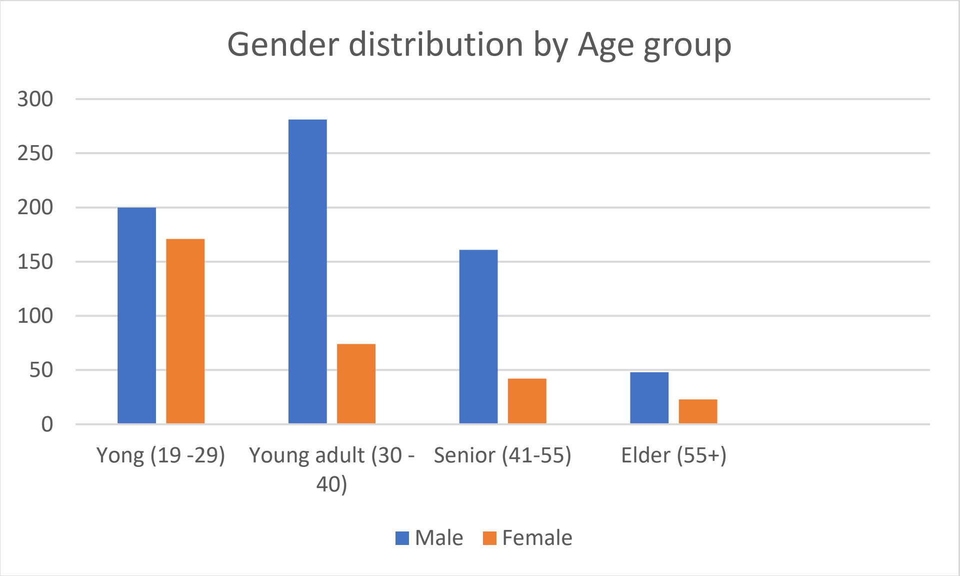


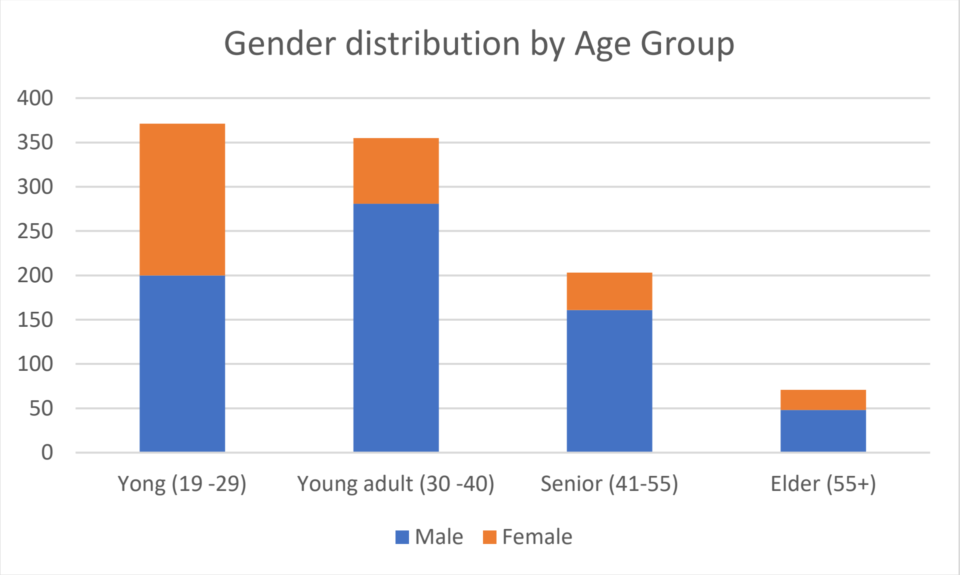
- Male applicants have better risk level than Female.

Age by gender.



	Male	Female
Yong (19 -29)	200	171
Young adult (30 -40)	281	74
Senior (41-55)	161	42
Elder (55+)	48	23

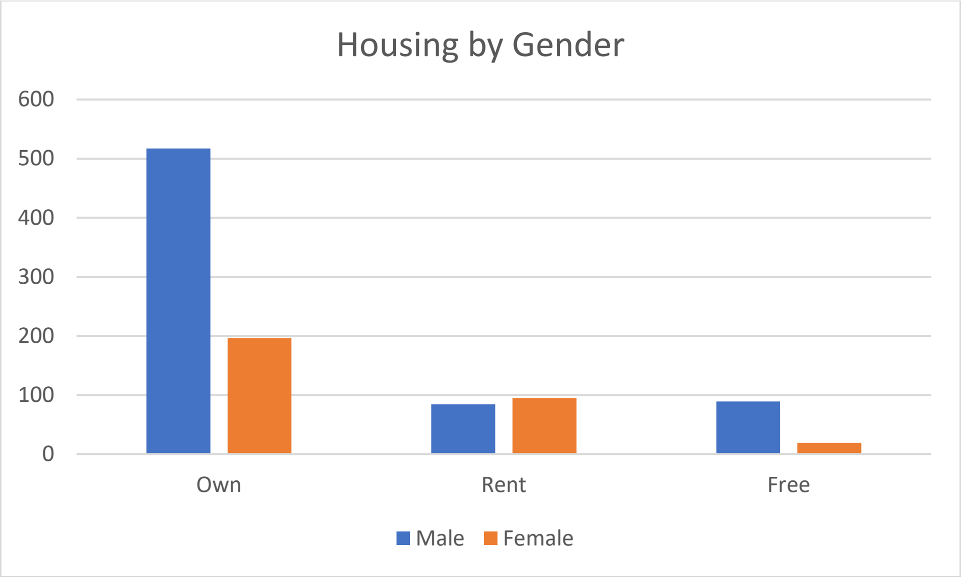


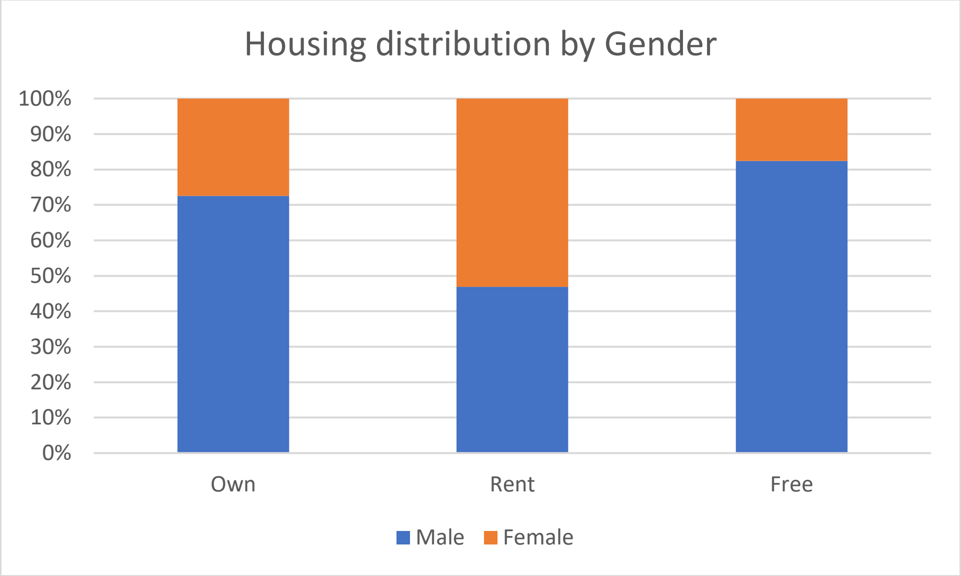


- At the young age (19-29) woman are applying for loan more likely than man.
- late in older ages woman are less likely to apply for loan.

Housing by Gender:

	Own	Rent	Free
Male	517	84	89
Female	196	95	19



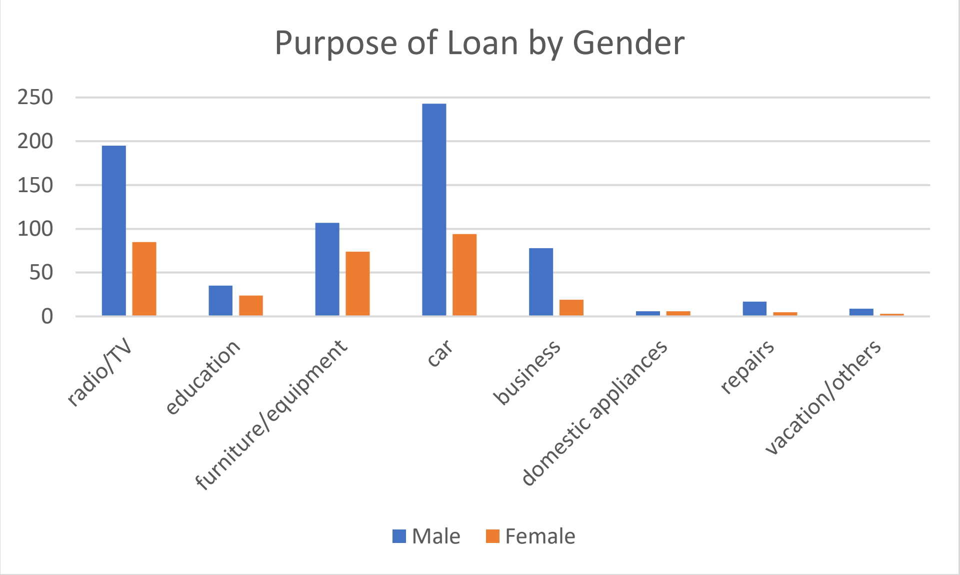


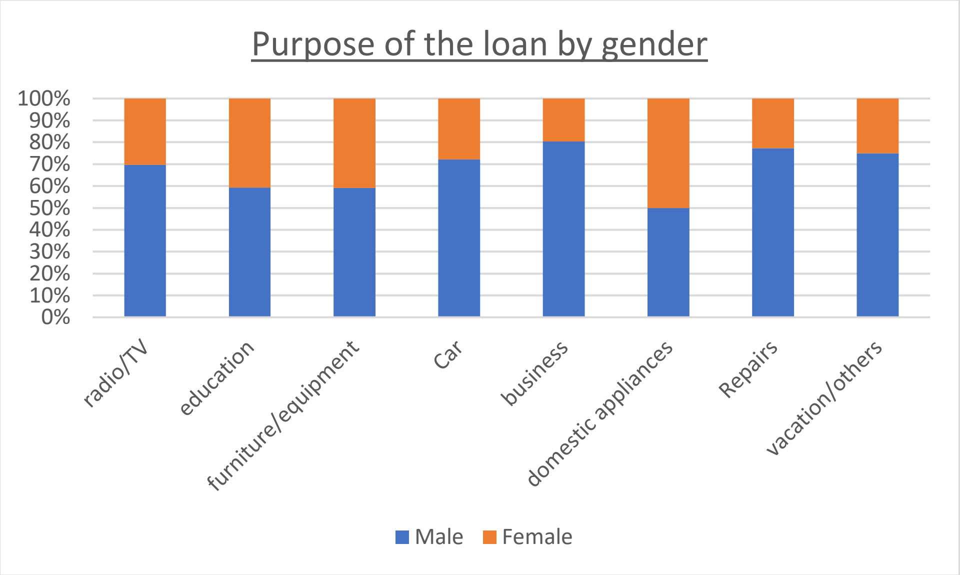
- Man has more own housing than female
- Man gets more free housing then female.
- In the case of renting housing female is slightly renting more than man.

Purpose of the loan by gender.

	Male	Female
radio/TV	195	85
education	35	24
furniture/equipment	107	74
Car	243	94
business	78	19
domestic appliances	6	6

Repairs	17	5
vacation/others	9	3



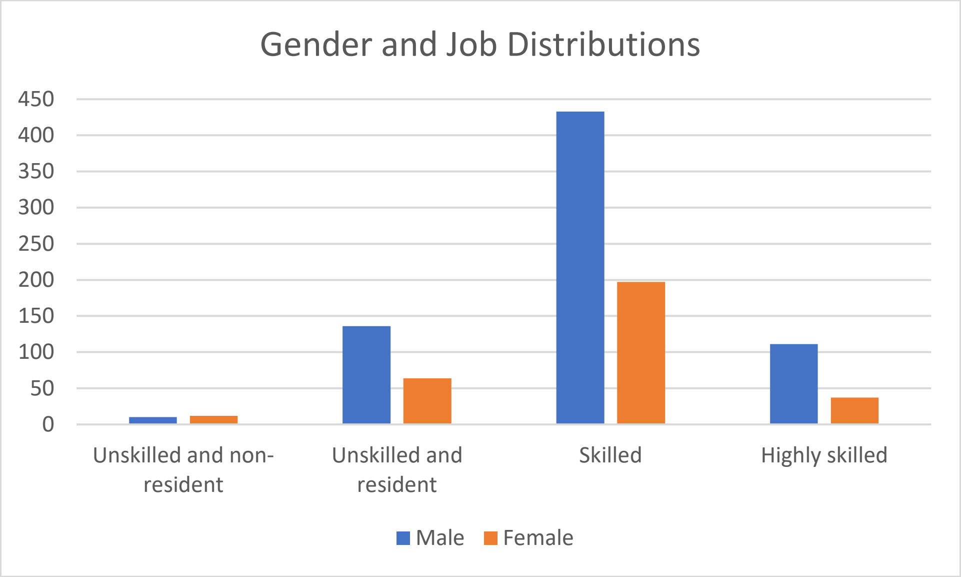


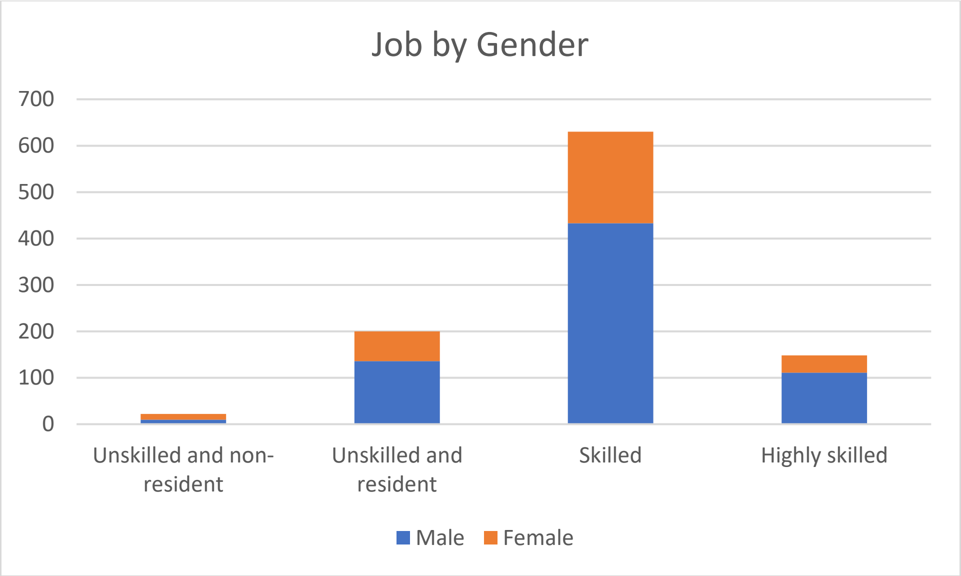
- Man, mostly likely to apply for more loan than woman on Business, Repairs, vacation and Car.
- In the case of domestic purpose Woman are equal to man for loan application.

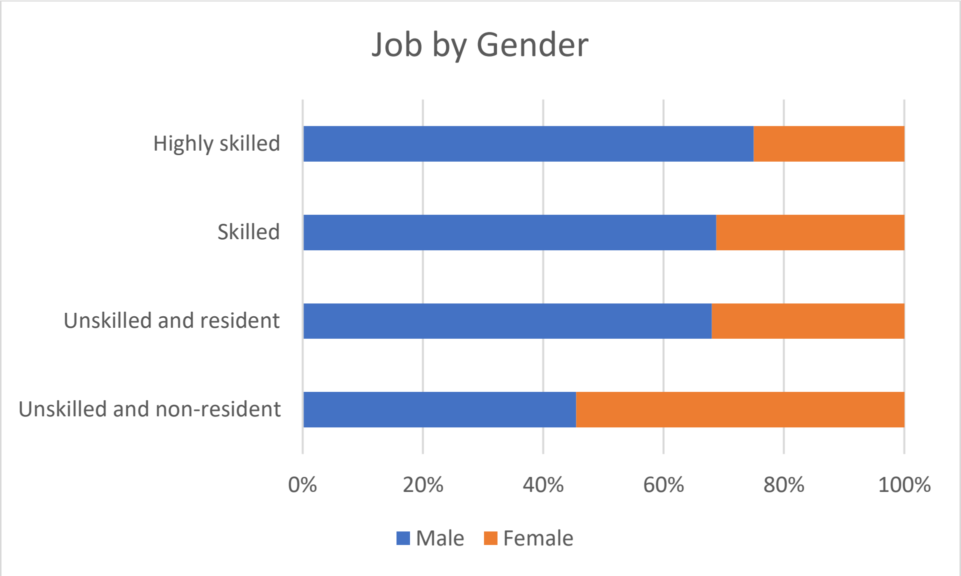
Employment status of applicants by gender.

	Male	Female
Unskilled and non-resident	10	12
Unskilled and resident	136	64
Skilled	433	197

Highly skilled	111	37

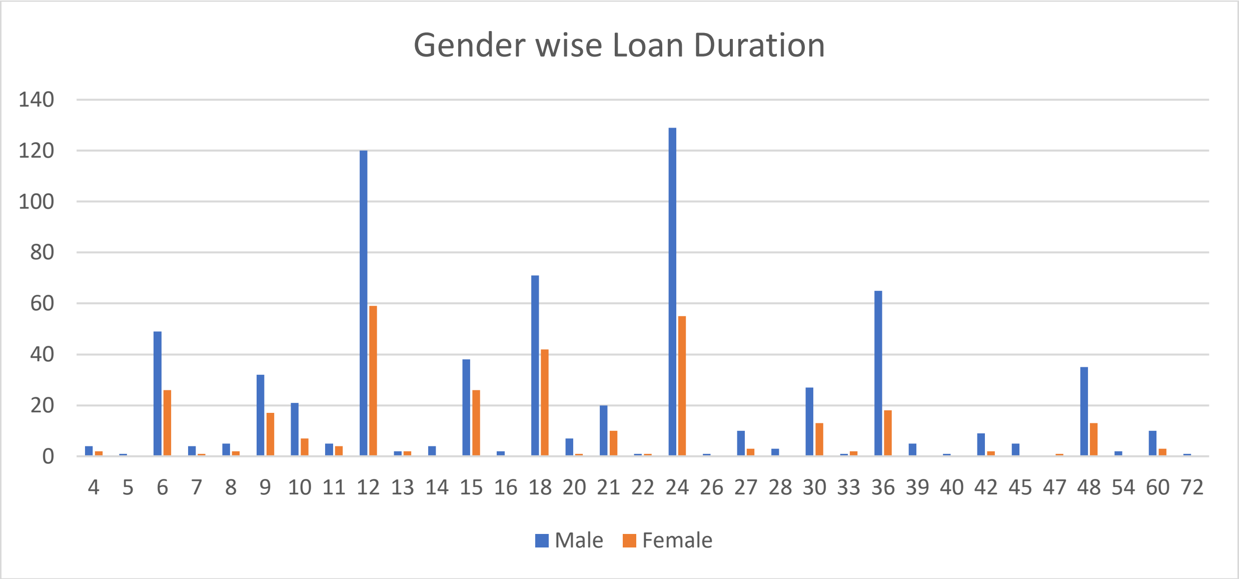


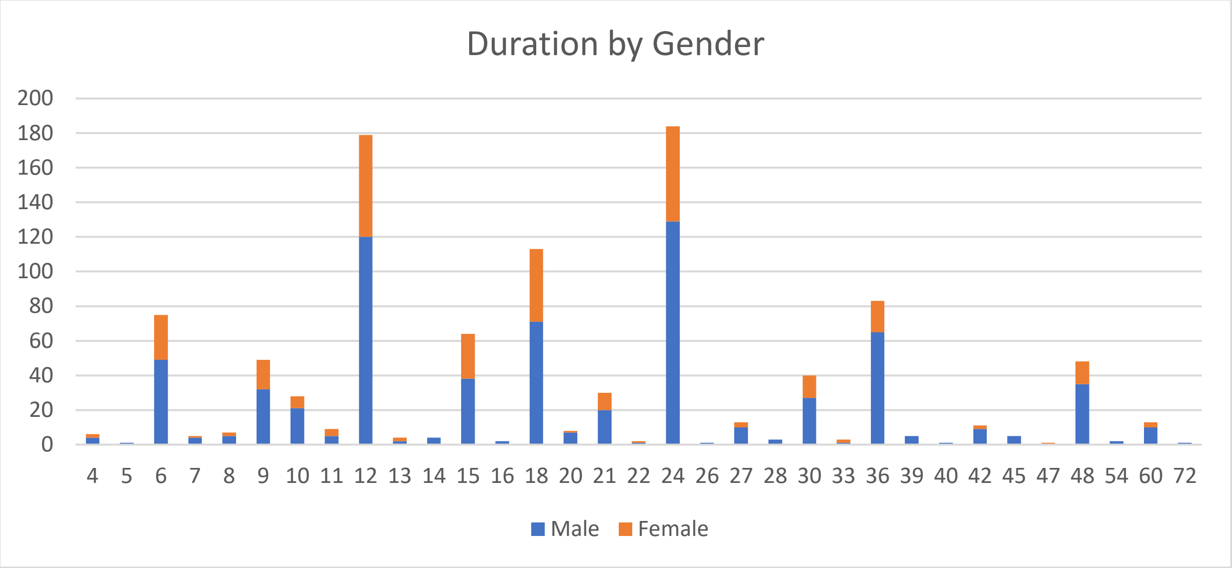




- Man is doing more skilled and highly skilled job than woman.
- In the case of unskilled and non-resident job group female are greater is number.

Duration by Gender:

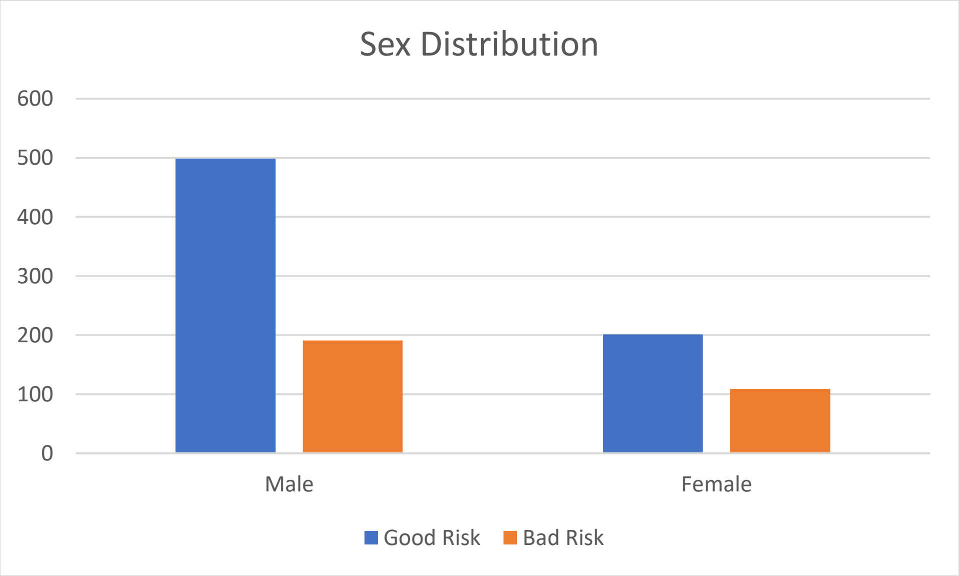


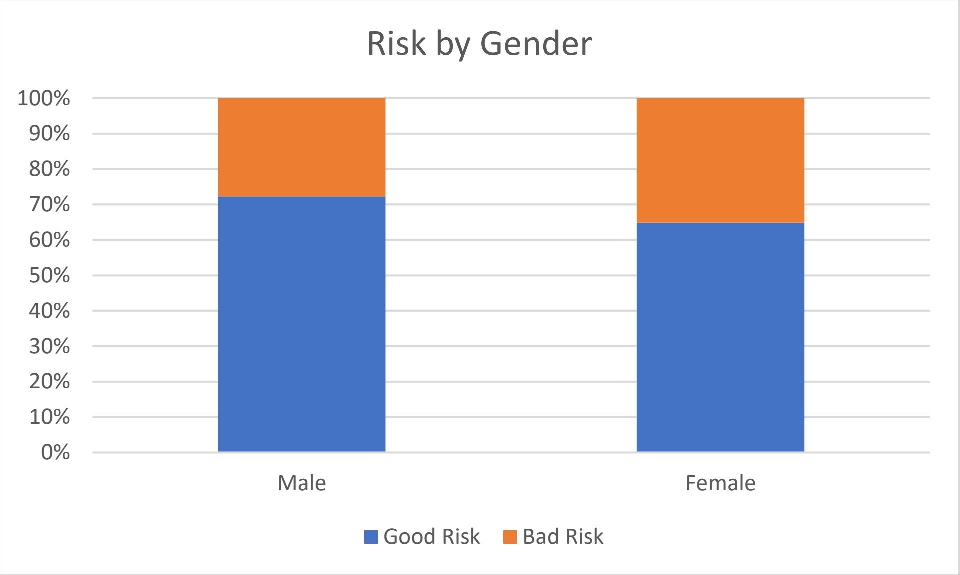


- Duration wise male and female are almost same.
- Both the gender (Male and Female) like to take short time loan which were mostly paid of by 24 Months.

Good loans and defaults (Risk) by gender:

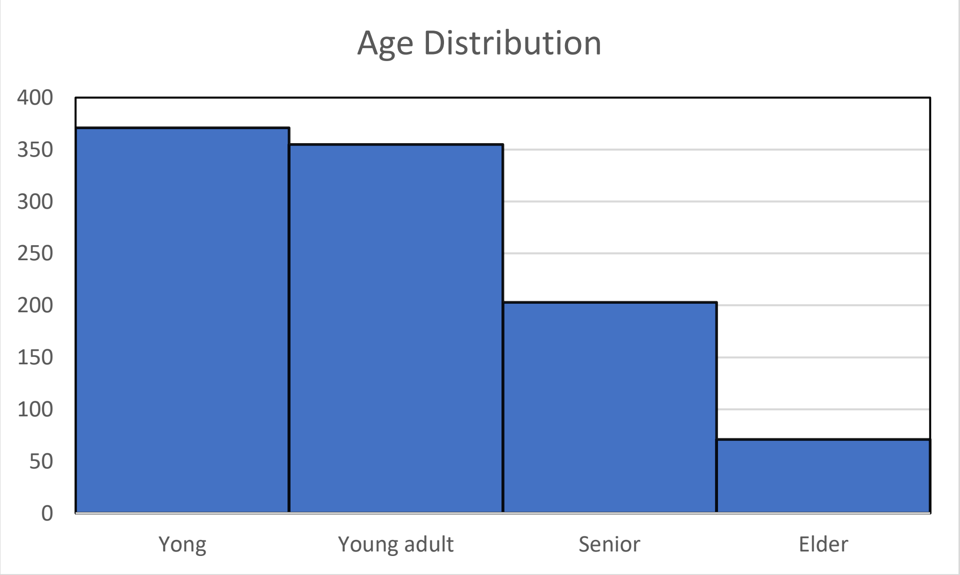
Risk	Male	Female
Good	499	201
Bad	191	109





- In loan risk, female are riskier than male.
- Around 72% Male has classified as good risk whereas around 65% Female has classified as good risk.

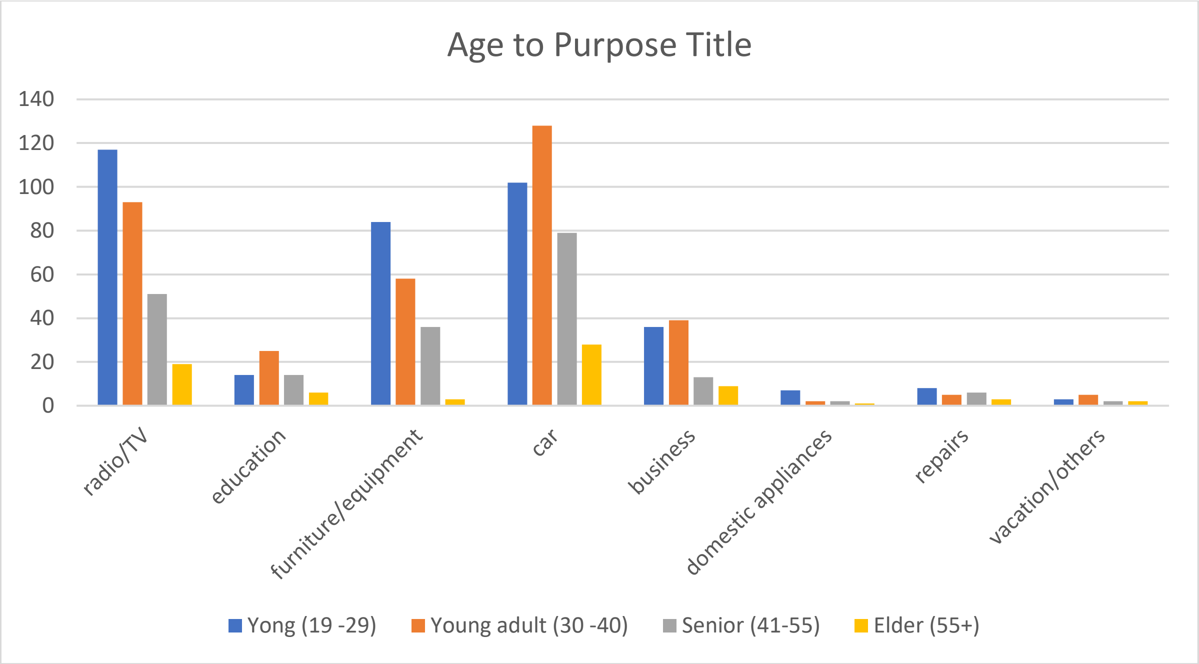
Bin	*Frequency*
Young (19-29)	371
Young adult (30- 40)	355
Senior (41-55)	203
Elder (55+)	71



- In the number of applications, Young people are making more loan application then the older people.

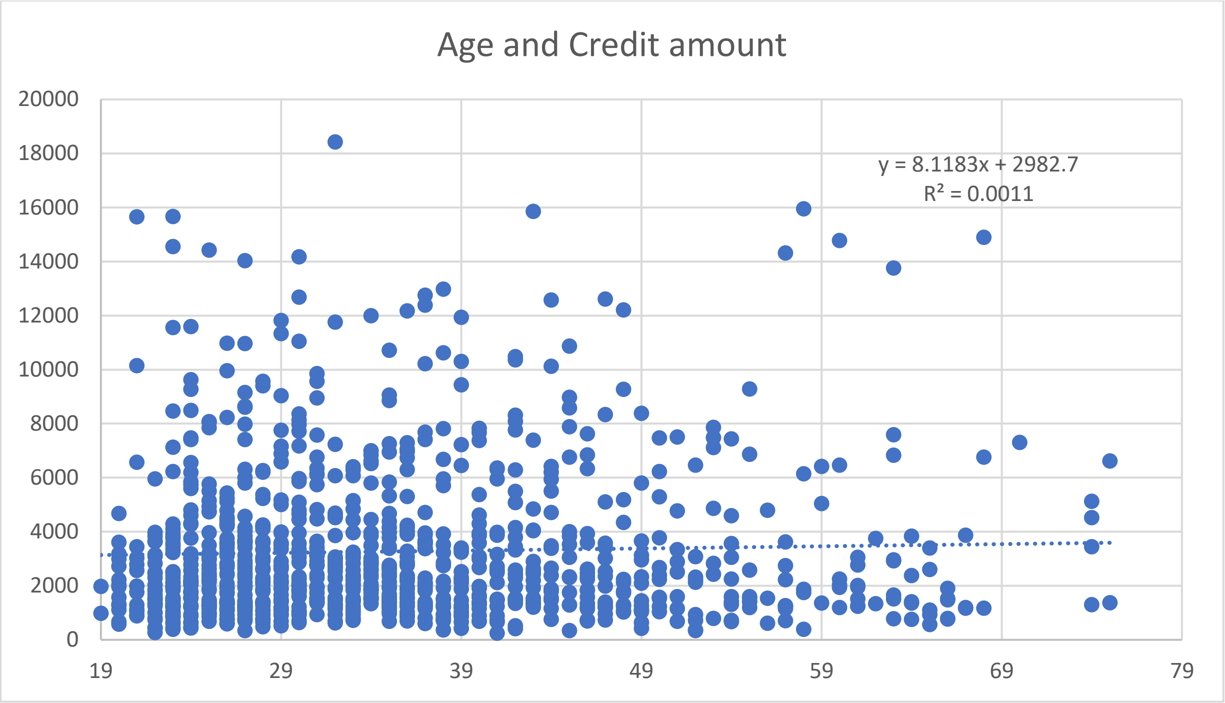
Age to purpose:

	radio/TV	education	furniture/equipment	car	business	domestic appliances	repairs	vacation/others
Yong (19 -29)	117	14	84	102	36	7	8	3
Young adult (30 -40)	93	25	58	128	39	2	5	5
Senior (41-55)	51	14	36	79	13	2	6	2
Elder (55+)	19	6	3	28	9	1	3	2



- Young (19-29) are more likely to apply loan for radio/TV.
- Young adult (30-40) are more likely to apply loan for Car.

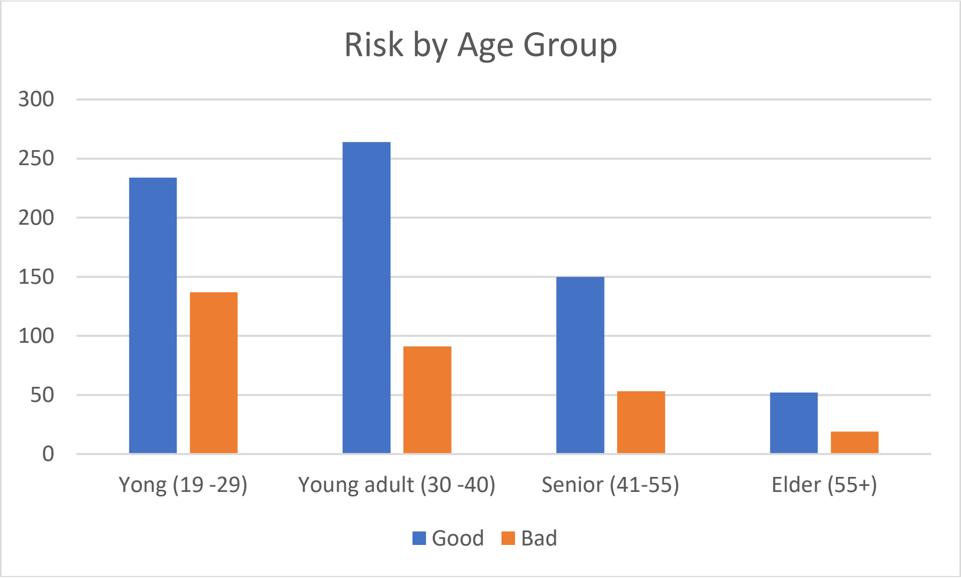
Age to loan size:

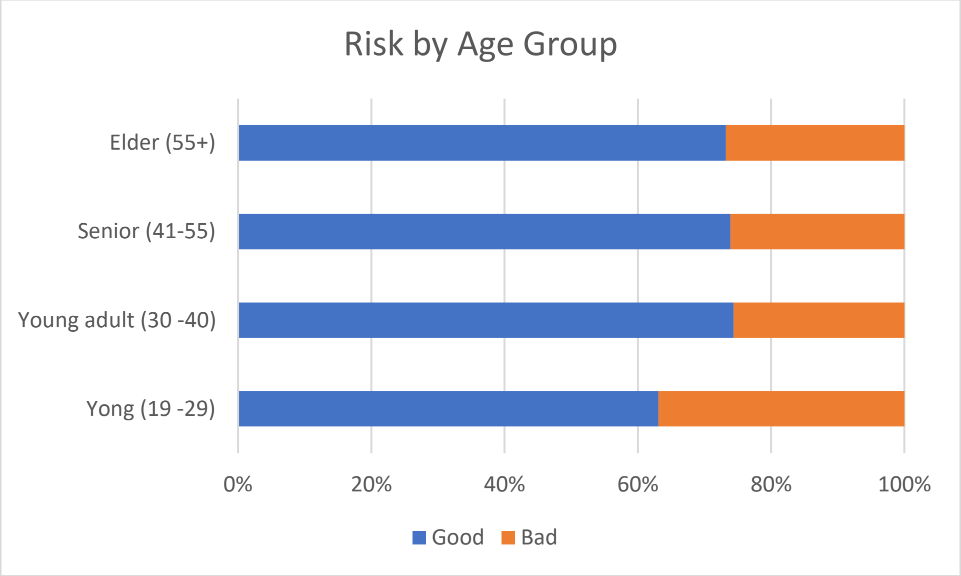


- There is slightly positive correlation between Age and Loan size.

Risk by age:

	Good	Bad
Young (19 -29)	234	137
Young adult (30 -40)	264	91
Senior (41-55)	150	53
Elder (55+)	52	19

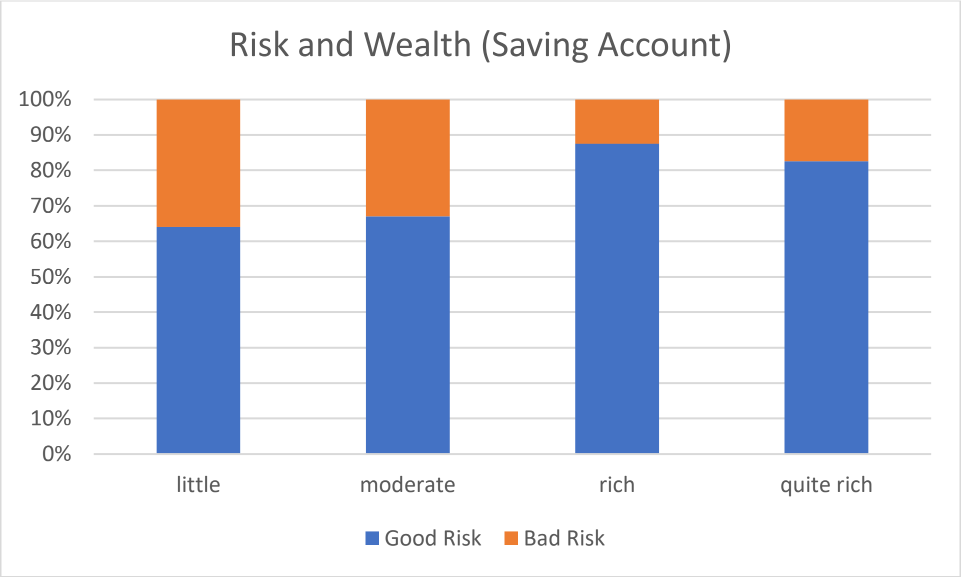




- Young (19-29) applicants are riskier than any other age group.
- Young (19-29) applicants are highest at Bad risk.

Risk by wealth:

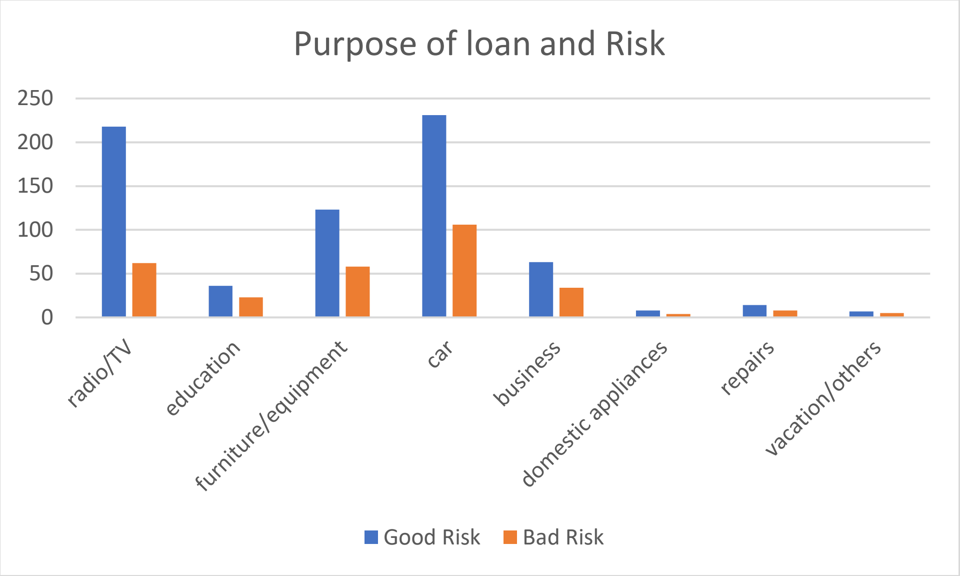
Risk	little	moderate	rich	quite rich
Good	386	69	42	52
Bad	217	34	6	11

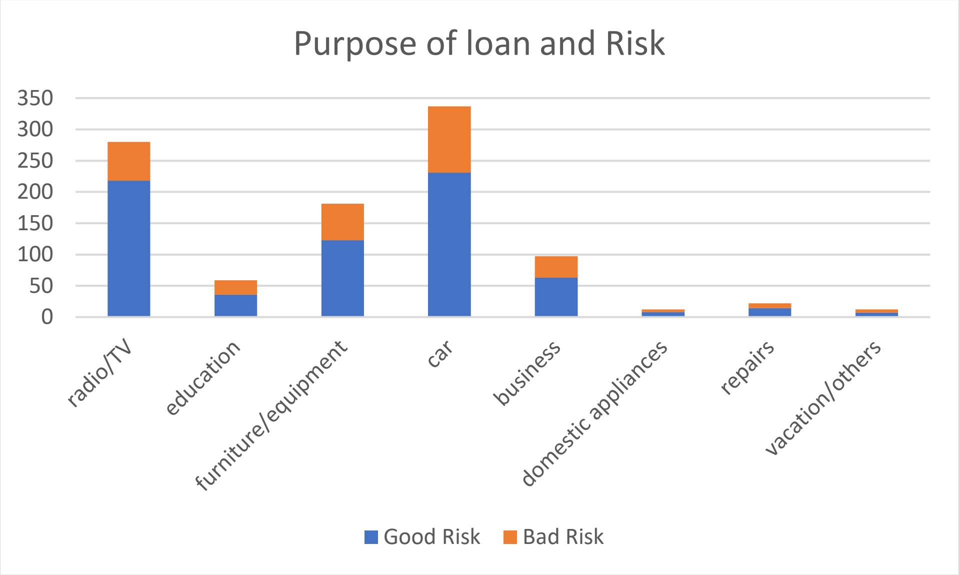


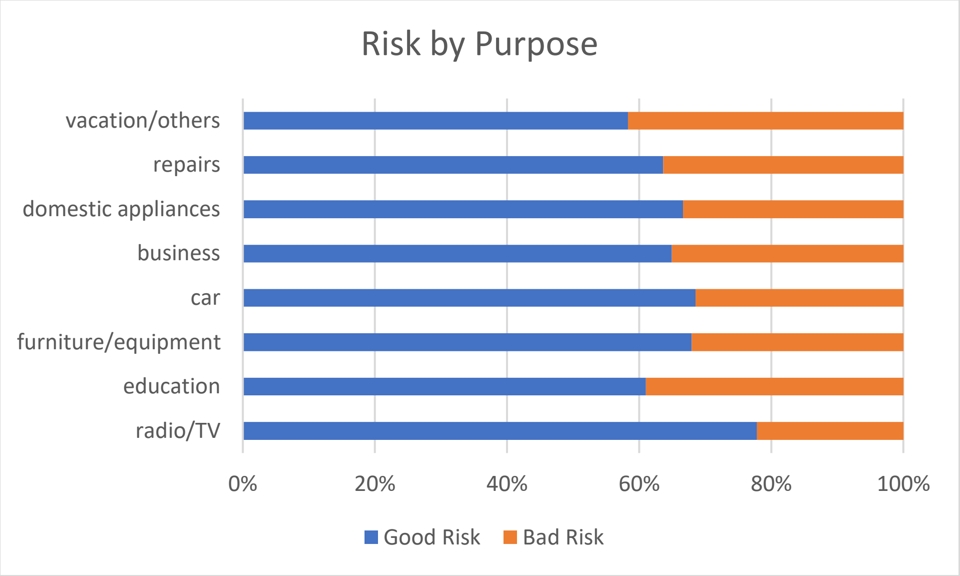
- Richer applicants are safer for loan.
- Richer are more likely more good risk.

Purpose of loan and Risk:

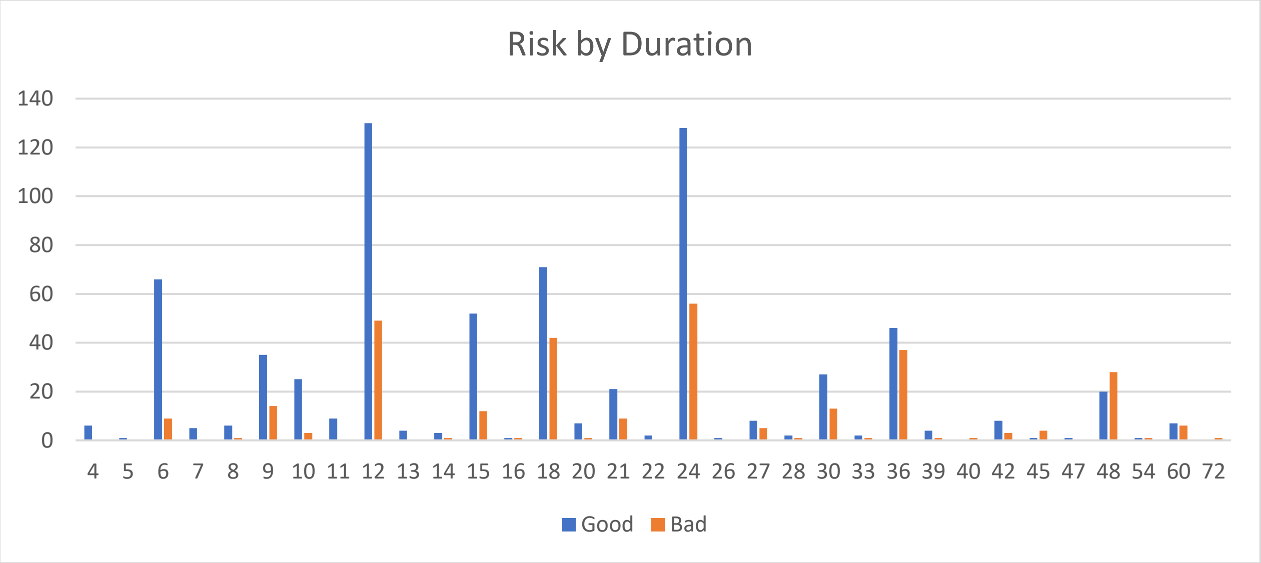
	Good Risk	Bad Risk
radio/TV	218	62
education	36	23
furniture/equipment	123	58
car	231	106
business	63	34
domestic appliances	8	4
repairs	14	8
vacation/others	7	5

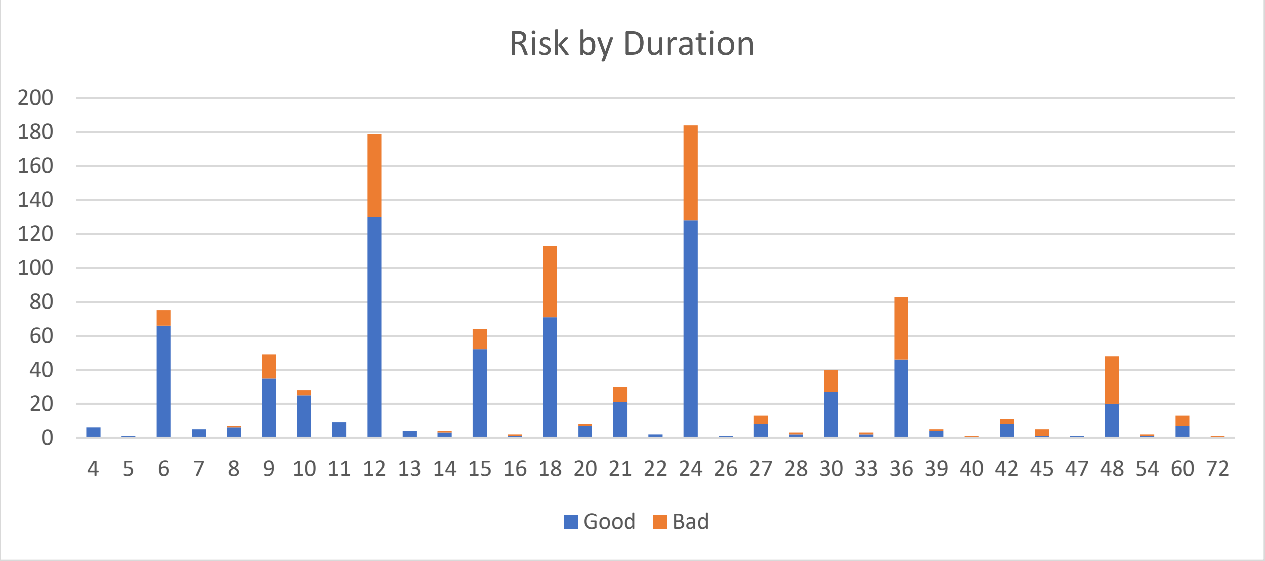

Purpose of loan and Risk
250
200
150
100
50
0
radio/TV
education
furniture/equipment
car
business
domestic appliances
repairs
vacation/others
Good Risk
Bad Risk

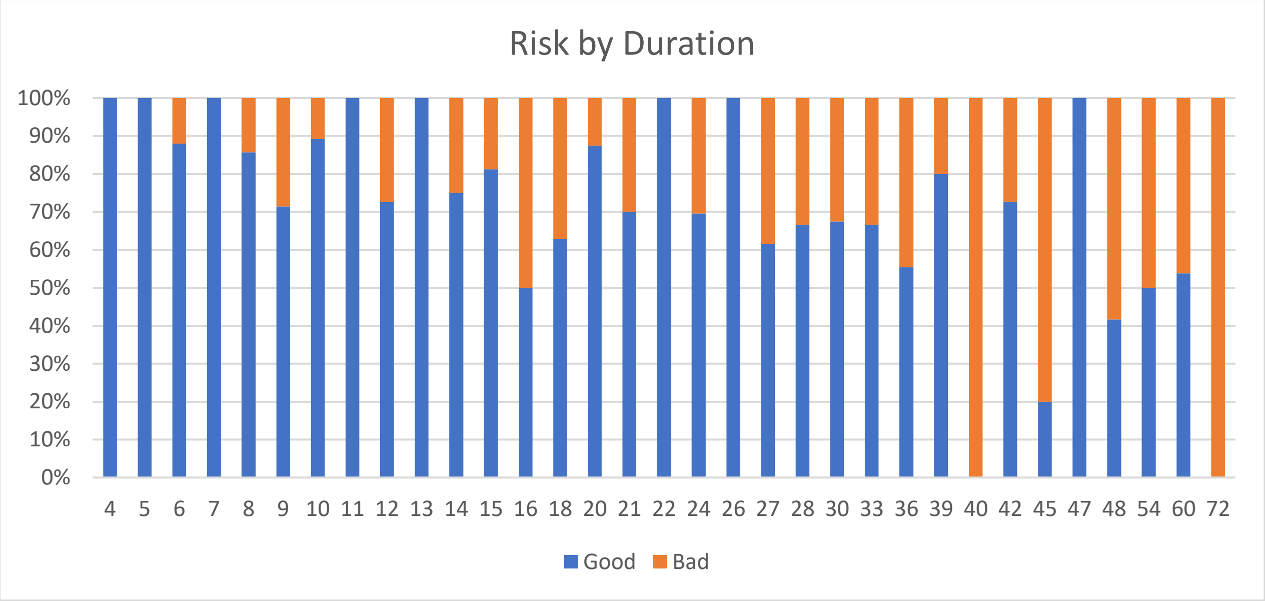

Purpose of loan and Risk
350
300
250
200
150
100
50
0
radio/TV
education
furniture/equipment
car
business
domestic appliances
repairs
vacation/others
Good Risk
Bad Risk



- The safest loan purpose is radio/TV
- The worst risk purpose is vacation/others.



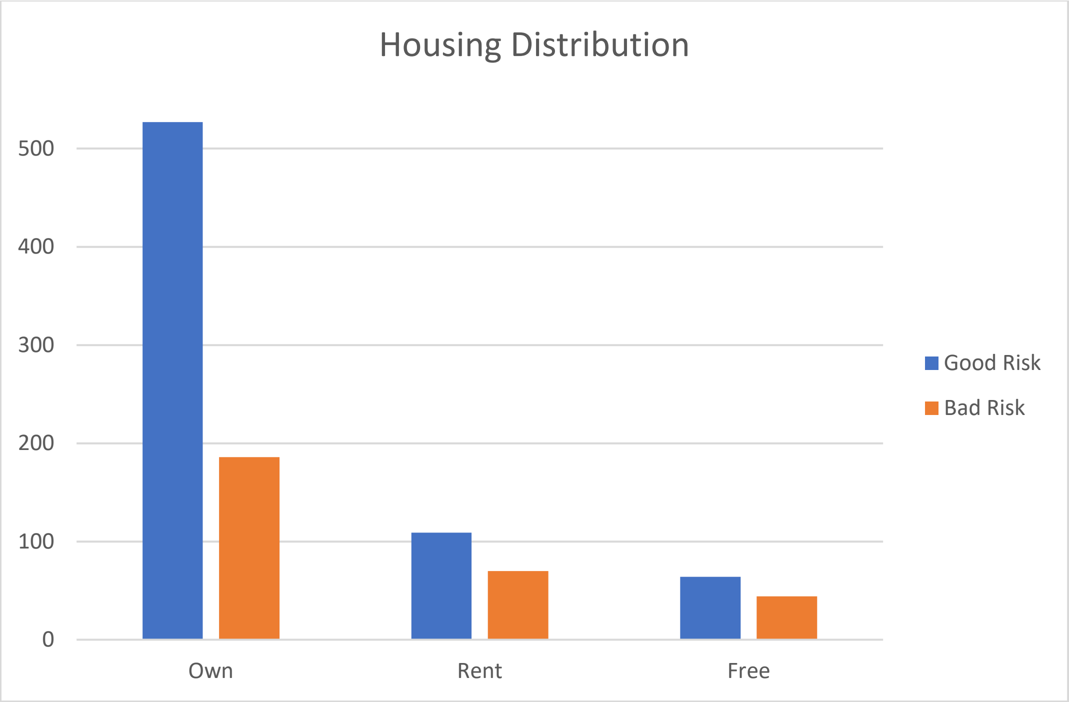


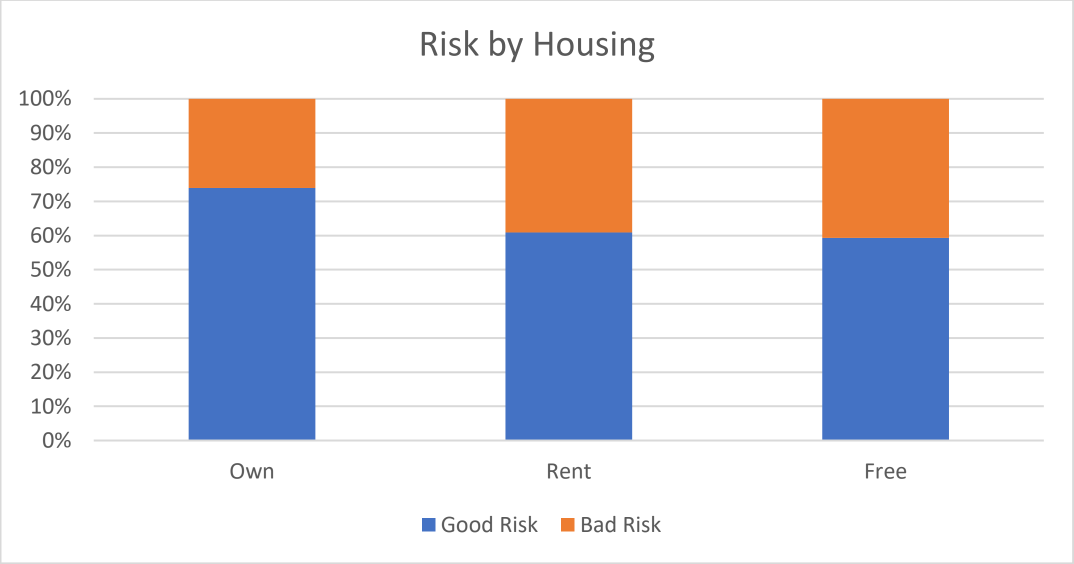


- Generally, loans for short duration are less risky than loan with longer duration.
- Shorter duration loans are good at risk.

Risk by Housing Distribution:

Risk	Own	Rent	Free
Good	527	109	64
Bad	186	70	44





- Loan Applicant with their own housing has the lowest bad risk.
- Over 70 % applicant with own housing is good risk.

Job and Risk:

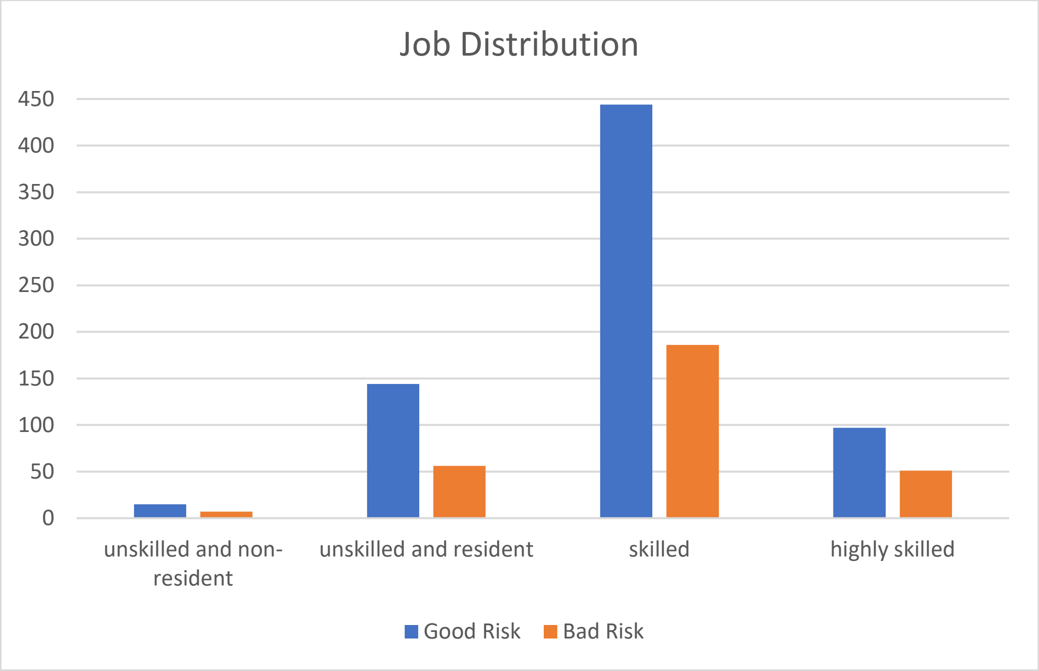
0 - unskilled and non-resident,

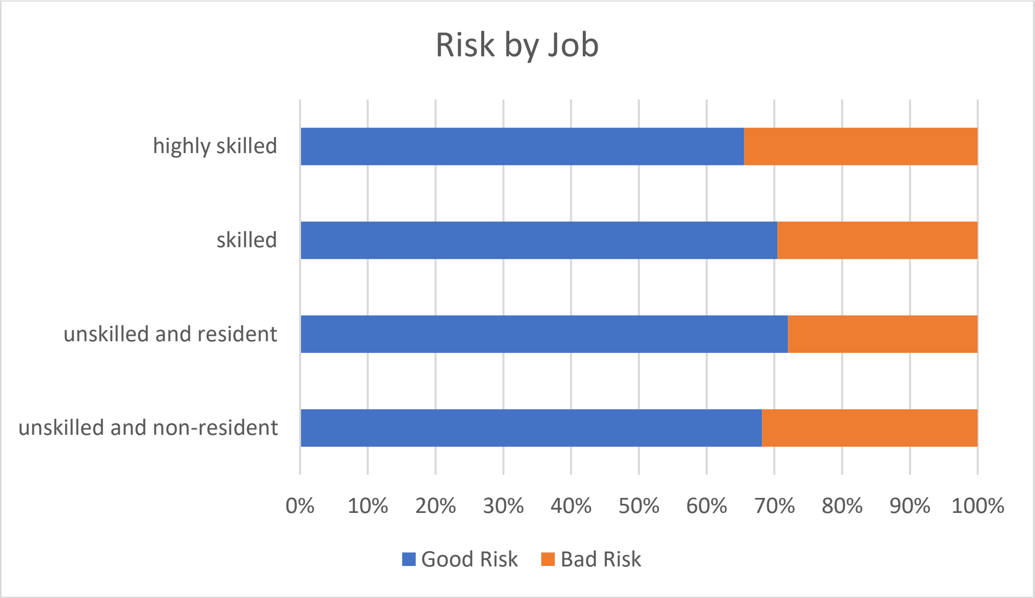
1 - unskilled and resident,

2 - skilled,

3 - highly skilled

Risk	unskilled & non-resident	Unskilled & resident	skilled	highly skilled
Good	15	144	444	97
Bad	7	56	186	51

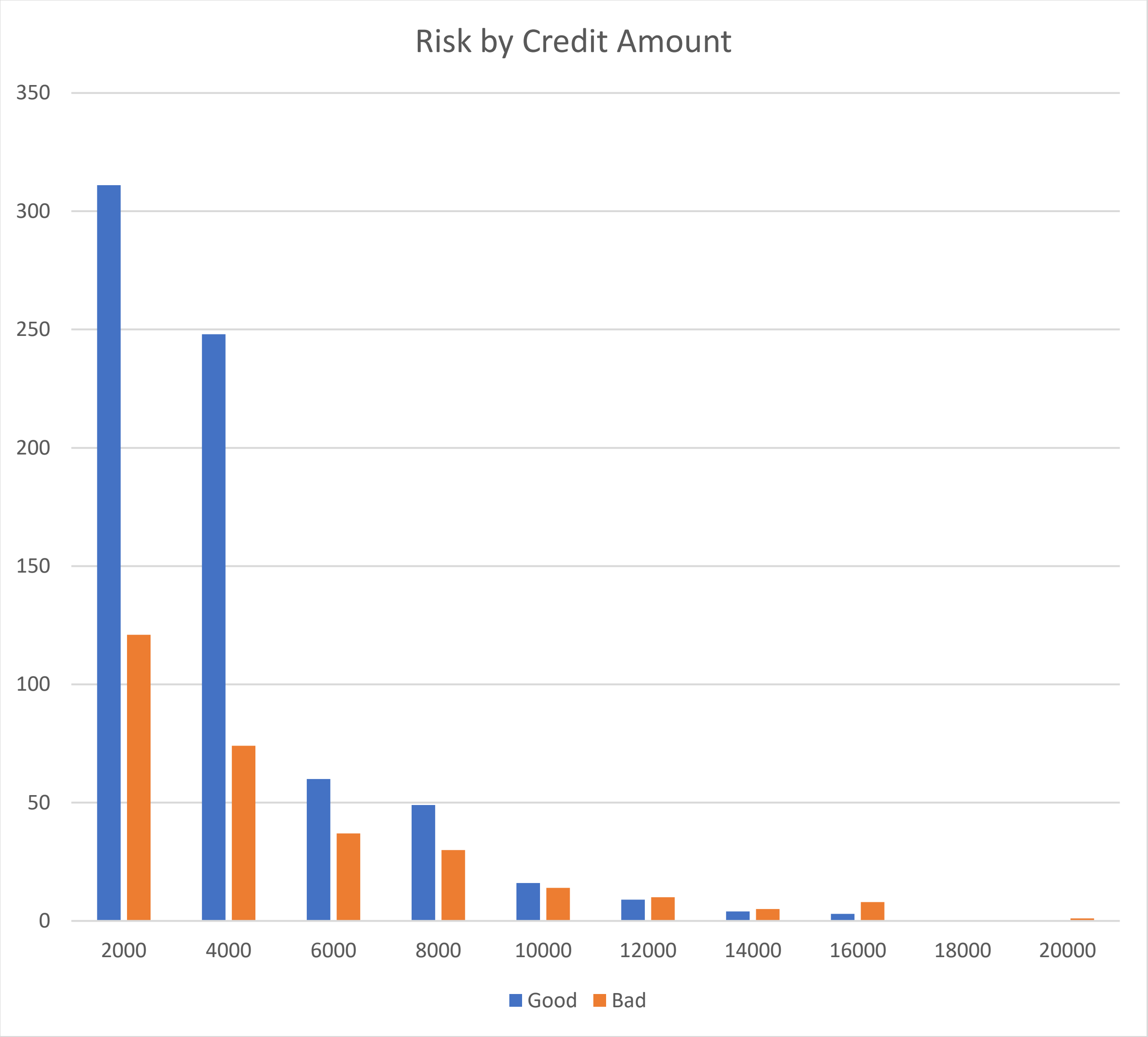


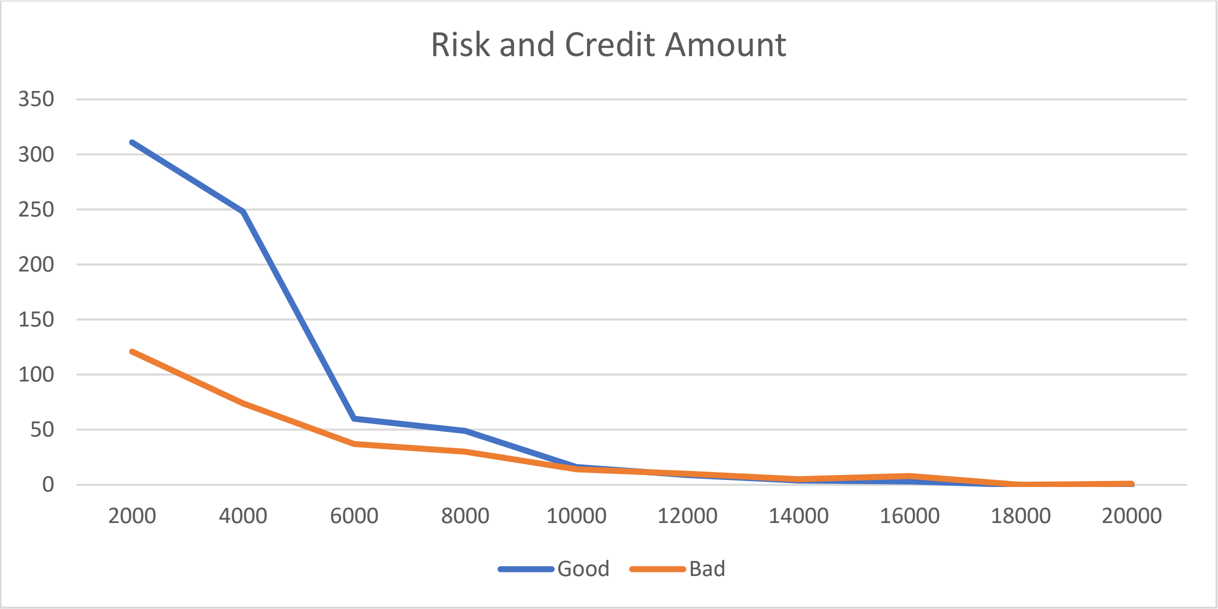


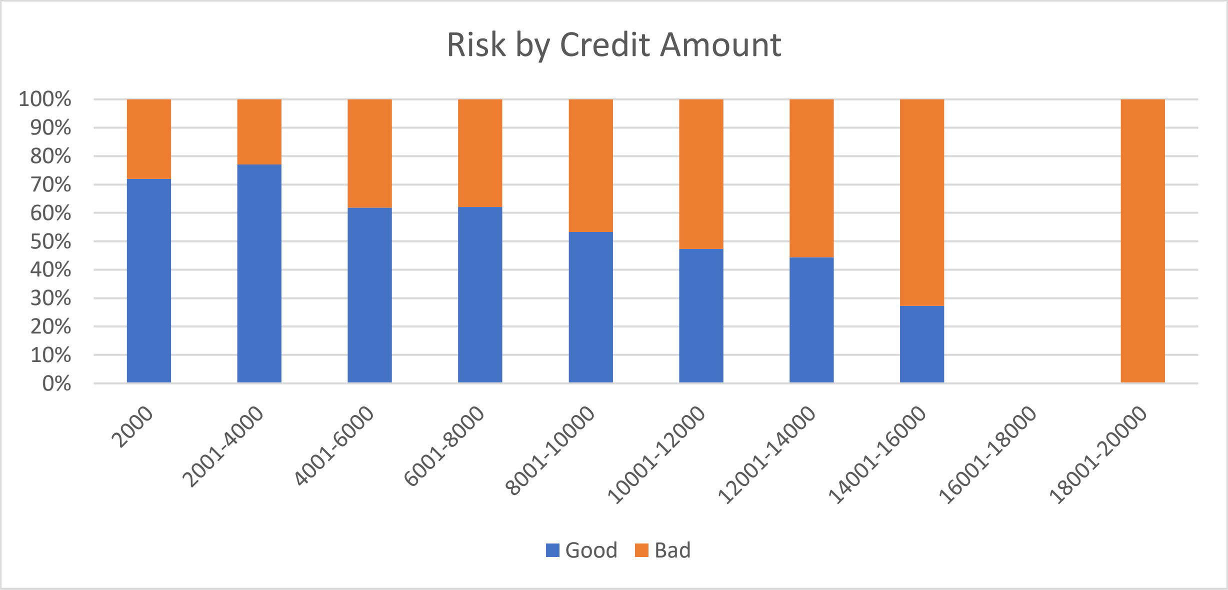
- In the case of loan, job status does not play huge role.
- In the case of lone, unskilled and resident applicant are comparatively better at risk. Thay have the lowest bad risk. Comparatively slightly higher bad risk is by the highly skilled job group.

Risk by Credit amount:

Credit amount	Good	Bad
2000	311	121
2001-4000	248	74
4001-6000	60	37
6001-8000	49	30
8001-10000	16	14
10001-12000	9	10
12001-14000	4	5
14001-16000	3	8
16001-18000	0	0
18001-20000	0	1


Risk by Credit Amount
350
300
250
200
150
100
50
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
Good
Bad

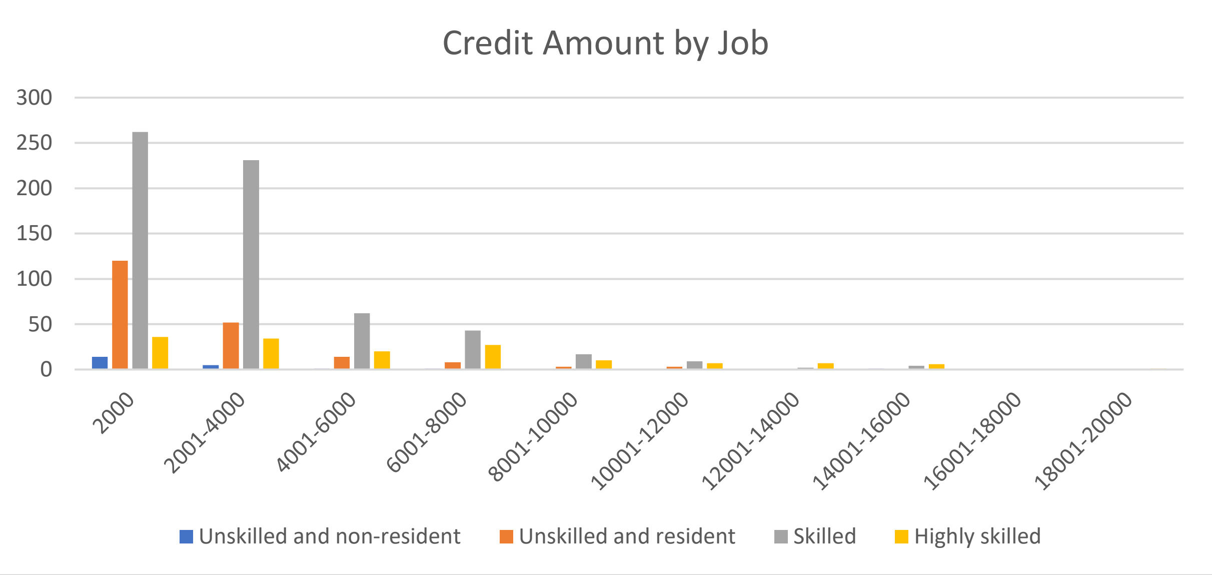




- Smaller Credit Amount loan has less bad risk than greater Credit Amount loan.
- Smaller Credit Amount loan has more likely classified as good risk, specially less than 4000 DM.
- Larger Credit Amount loan has more likely classified as bad risk.

Credit amount	Unskilled and non-resident	Unskilled and resident	Skilled	Highly skilled
2000	14	120	262	36
2001-4000	5	52	231	34
4001-6000	1	14	62	20

6001-8000	1	8	43	27
8001-10000	0	3	17	10
10001-12000	0	3	9	7
12001-14000	0	0	2	7
14001-16000	1	0	4	6
16001-18000	0	0	0	0
18001-20000	0	0	0	1

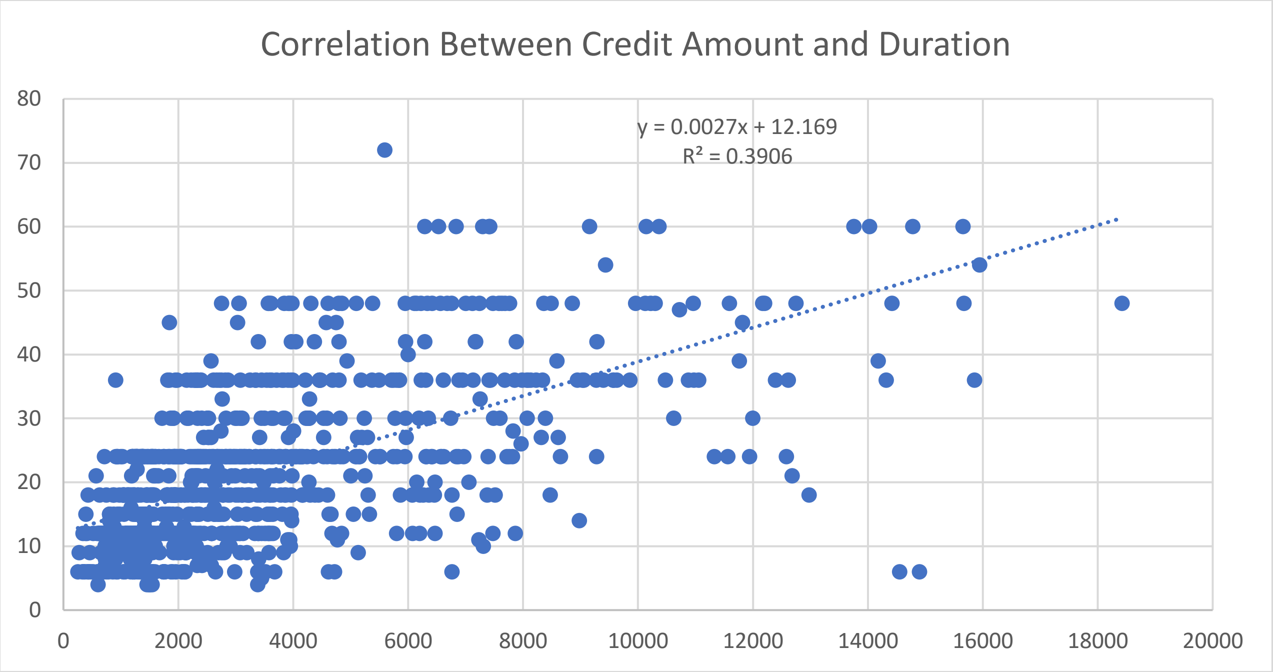


- Skilled Job group are applying for more loan than any other group.
- Skilled and highly skilled job group are more likely to apply big amount of loan.

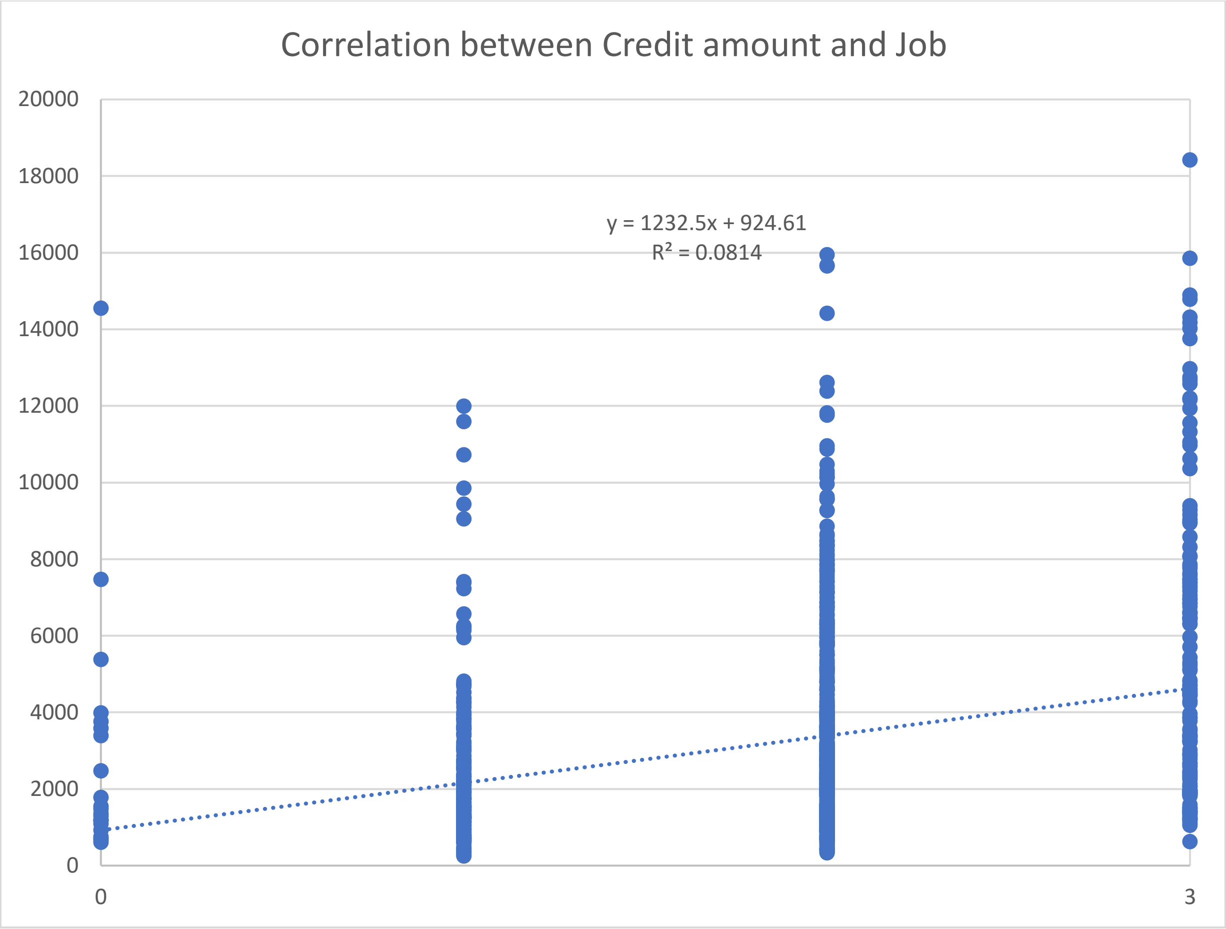
Correlation matrix:

It tells about the relationship between two features (variables). The relationship means a linear correlation. The value ranges between -1 to +1. Positive value means if one variable increases the other variable also increases. On the contrary, a negative value means if one variable increases the other variable decreases. 0 value means there is no relationships between the variables.

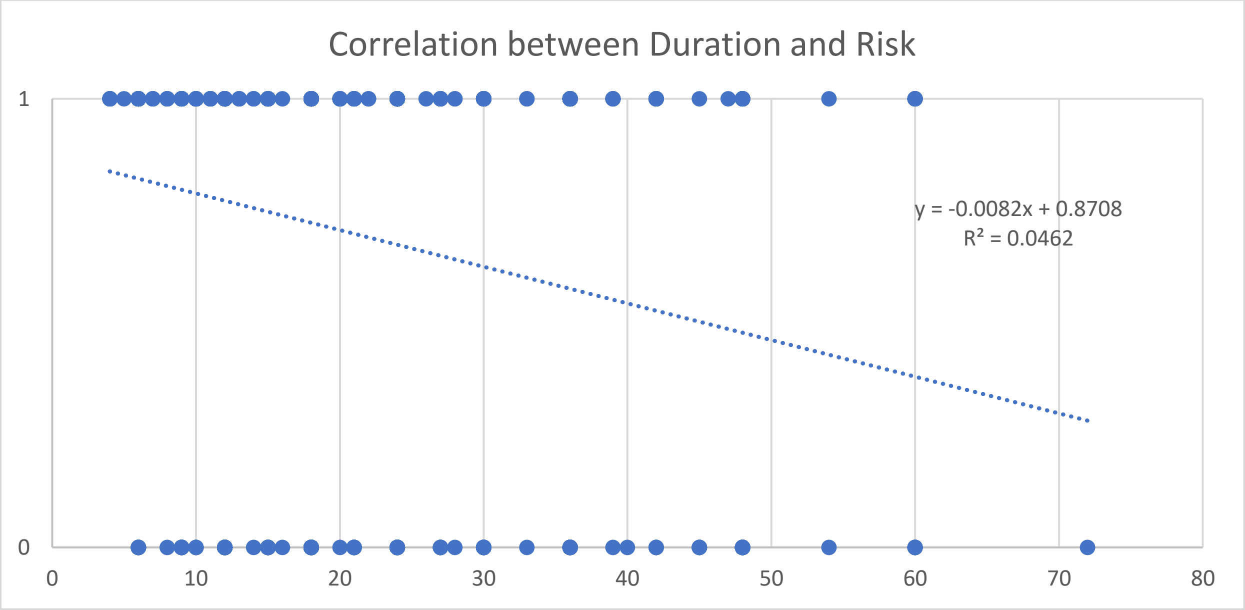
	Age	*Job*	*Credit amount*	*Duration*	*Risk int*
Age	1				
Job	0.015673	1			
Credit amount	0.032716	0.285385	1		
Duration	-0.03614	0.21091	0.624984198	1	
Risk int	0.091127	-0.03274	- 0.154738641	- 0.214926665	1



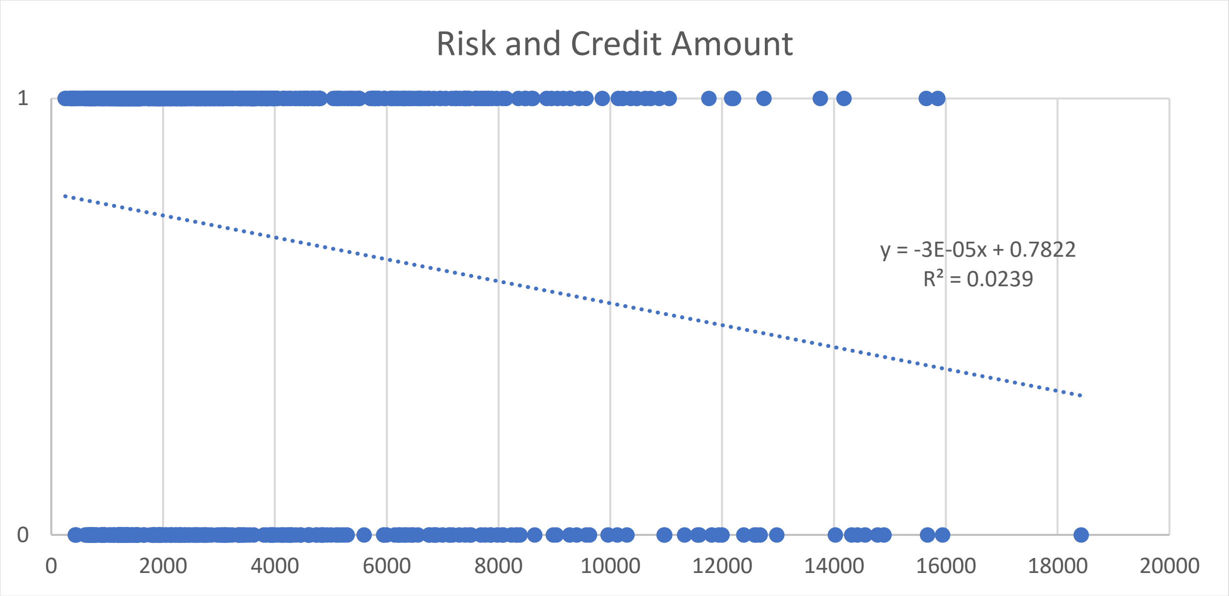
- A Positive Correlation Between Credit Amount and Duration with 0.62.
- The variables are moving to the same direction.



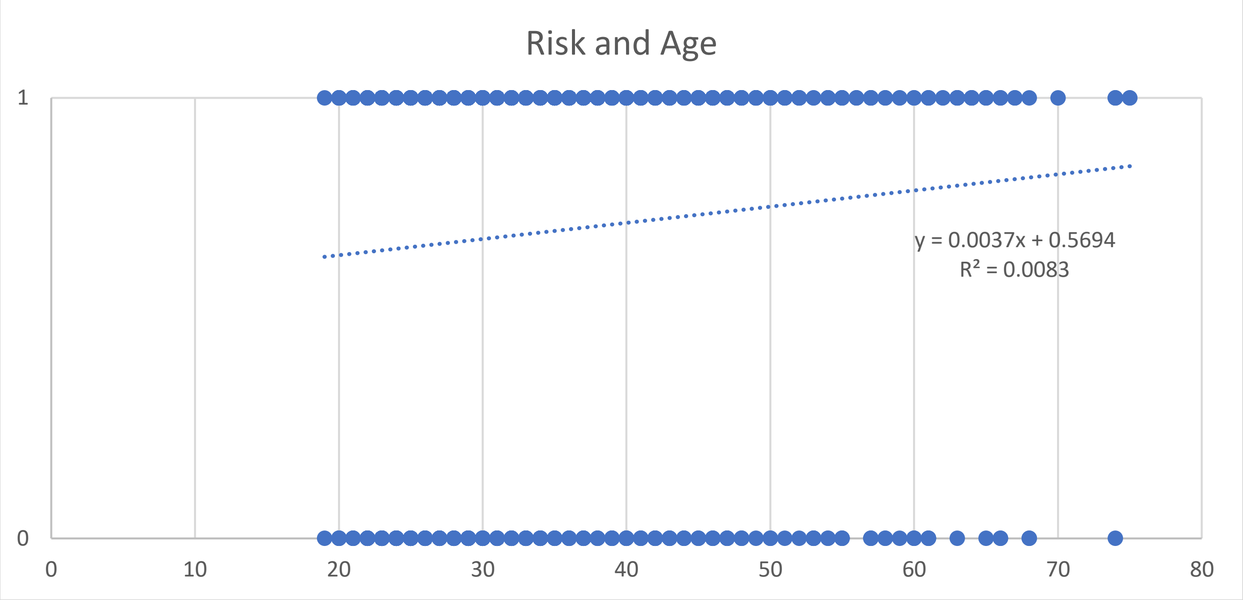
- A positive Correlation between Credit amount and Job with value 0.28



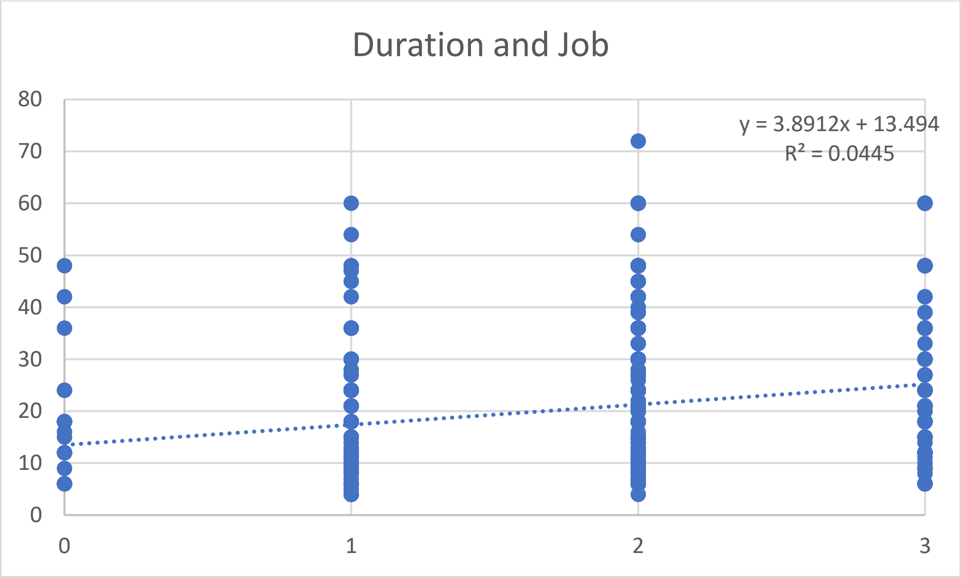
- A Negative Correlation Between Risk and Duration with -0.21



- A Negative Correlation Between Risk and Cradit Amount with -0.15



- A Positive Correlation Between Risk and Age with 0.09



- A Positive Correlation Between Job and Duration with 0.21